VOLUME 52
No. 3

WHOLE NO. 234
1940

Psychological Monographs

EDITED BY
JOHN F. DASHIELL
UNIVERSITY OF NORTH CAROLINA

Studies in Quantitative Psychology

From the University of Illinois

Edited by
HERBERT WOODROW

PUBLISHED BY
THE AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.
PUBLICATIONS OFFICE
THE OHIO STATE UNIVERSITY, COLUMBUS, OHIO

TABLE OF CONTENTS

		PAGE
1. The Problem of the Interrelationship of Determining Conditions	HERBERT WOODROW	1
2. The Effect of a Fixed Change in Difficulty at Various Levels of Difficulty	JOHN M. WILLMANN	9
3. The Measurement of Memory on an Absolute Scale	HARRIETT C. SHURRAGER	21
4. A Factor Analysis of Forty Character Tests	HUBERT E. BROGDEN	39
5. The Effect of Practice upon Standard Errors of Estimate	LELAND P. BRADFORD	56

THE PROBLEM OF THE INTERRELATIONSHIP OF DETERMINING CONDITIONS

By

HERBERT WOODROW

Two of the following investigations, those by H. Shurrager and by Willmann, deal with a problem which the writer has been studying for a number of years, that of how the different determining conditions of goodness of performance combine. It may be helpful, as an introduction to these two studies, briefly to outline the general problem towards the solution of which they contribute.

As applied to task performances, probably all would agree that the general hypothesis of determinism could be expressed by the following equation:

$$S = f(x_1, x_2, x_3, etc. \ldots . x_n) \quad . \quad . \quad . \quad (1)$$

in which S represents the goodness of score and x_1, x_2, x_3, etc. . . x_n are its determining conditions. The interrelation between these conditions is left open by a formula written in the above fashion.

One of the above terms, namely, goodness of score, may demand explanation. All scores of the sort here considered are scores made with respect to some abstract attribute of performance or its immediate products, such as speed, time, percentage correct, percentage of responses of a certain category, number of errors, etc. It is a matter of common consent that some scores made on any scale measuring such attributes are better than others. Thus, to react quickly is better than to react slowly (if the instruction is to react as quickly as possible); to judge a small difference correctly in 40 per cent of the trials is better than to judge with equal correctness only a larger difference; and to judge a given difference correctly in 60 per cent of the trials is better than to judge it correctly in only 40 per cent of the trials. Scores at one end of certain scales are thus considered to be

relatively good and those at the other end relatively bad. As here employed, "goodness" is a term applying to both the good and bad scores, so that the former merely have more goodness than the latter. This usage is similar to that often found in the case of intelligence, speed, accuracy, etc. It is simpler to regard intelligent and unintelligent, slow and fast, accurate and inaccurate, as each differing only with respect to one attribute.

It is not altogether easy to say why one score is better than another. In general, it is no doubt true that by goodness of score is understood the efficiency with which the subject carries out some purpose or attains some end, either on his own initiative or because of the experimenter's instructions. It seems preferable, however, to define goodness in a more precise manner. There are perhaps several satisfactory ways of doing this. One way is to say that the performance high in goodness is one of which relatively few people are capable. From this viewpoint, it may be said that it is better to solve all ten problems of a test than to solve only three, because fewer people can solve ten than three. But even if we are considering only a single subject, it is still possible to distinguish the good performance from the poor. This may be done by observing that to make a score of a certain goodness a person must also be able to make the less good scores. Thus, a person who can remember eight or ten nonsense syllables obviously can also remember two. The converse is not true. Likewise, the child who passes intelligence tests so as to obtain a mental age of eight can also pass tests (and will, if they are given him) sufficient to obtain a score of five. The child whose mental age is scored as five, however, cannot pass the tests required to obtain a mental age of eight. It is thus possible to distinguish the good end of a scale from the poor end without any reference whatsoever to the usefulness of the performance or to social environment.

Performances giving scores which in this last sense may be regarded as varying in goodness may be termed task-performances. Not all instances of behavior have this attribute. Apparently many "personality" traits lack the attribute of goodness. It would not make sense, for example, to say that the person

making a high introvert score could also make a low one. The present discussion is limited to task-performances which possess the attribute of goodness.

Now any equation showing the relation between the goodness of performance and its determining conditions will depend upon the kind of units in terms of which goodness is measured. The magnitude of the unit is relatively inconsequential and arbitrary. The type of unit, however, is enormously important. Consider, for example, simple reaction-time to sound. It is of no particular importance what time-unit one uses, but it is necessary to decide how goodness varies with the time. Obviously the relation is inverse, since the reaction giving the smallest reading is the best. Shall one, however, measure goodness as $k-t$ (t standing for time); or as $\frac{1}{t}$; or as $\frac{1}{log\ t}$; or as some other function of t? Clearly any equation expressing the relation between goodness of reaction-time and its determining conditions depends upon the answer to this question. And any general properties of such equations, that is, characteristics which might apply to a whole set of equations each for a different performance, in other words, any general laws governing the relation between goodness of performances and their determining conditions, would certainly not be discovered if the question of how to measure goodness of response were improperly answered.

One of the most important characteristics of an adequate unit of measurement is undoubtedly that it remain constant at all parts of the scale. Any scale for measuring goodness of performance in equal units is here termed an absolute scale, even though the zero value be unknown. Any technique which results in such a scale may be termed an absolute scaling technique. The technique of absolute scaling is still in its infancy. All existing techniques are based on assumptions (such as that of a normal distribution of ability in a certain population) that have not been, and perhaps cannot be, directly proved or disproved. Apparently the validity of the necessary assumptions must for the present rest upon the degree to which their use leads to worthwhile outcomes.

As the first step in an attempt to work out equation (1), measurements have been made of goodness of performance in scaled units, so that, in further study, equation (1) may be considered as it applies to the score, S, when S is in terms of an equal unit scale. In the case of reaction-time, the conclusion was reached (2) that goodness (measured in equal units) varies as $log\left(\frac{1}{t}\right)$, t being the reaction-time; while in the case of several per cent correct scores (for example, in the following study of recall of nonsense syllables), the relation between raw score and goodness score is expressed by the ogive of a normal distribution curve.

The problem of the interrelationship of determining conditions of goodness scores has been studied by the writer chiefly by attempting to determine whether the equation for such scores could be written, more definitely than above (equation 1), in the following form:

$$S=f_1(x_1)+f_2(x_2)+etc....+f_n(x_n) \quad . \quad . \quad . \quad (2)$$

This equation was not suggested by *a priori* considerations, nor by the fact that somewhat similar equations are used in factor analysis to express the dependence of standard score upon the various determining factors, the latter being ordinarily conceived to be abilities. On the contrary, the equation was plainly indicated by the data of the first experimental investigation of the matter (1). Equation (2) carries no implication concerning the nature of the relationship between scaled goodness score, S, and any one determining condition. Only in a few instances has an attempt been made to determine the nature of the functions indicated by the f's of the equation, but it seems probable that the relation between S and any determining condition is seldom linear. For example, an empirical equation worked out for the relation between S, when it consisted of the goodness of performance in the case of naming two letters exposed for .1 sec., and the spatial separation of the letters, s, was the following:

$$S=a+b\ log\ (s+k)$$

in which a, b, and k are constants. If one wished to do so, one

could no doubt from the data here presented by Shurrager write the equation expressing the relation between goodness of score and length of series in the case of recalling nonsense syllables.

It is unnecessary, however, to know these functional relationships between S and each of its determining conditions in order to investigate the above formula. For whatever form $f(x)$ may take, the difference in its value in the case of two different particular settings of x will be a fixed value. Such a change in some one condition from one particular constant value to another particular constant value is termed a fixed change in conditions. The whole problem here under investigation may be answered merely by determining whether the change in the value of S due to a fixed change in x_1 remains constant when various values are given to x_2, x_3, etc. If the effect of the fixed change is found to be constant, that result obviously supports the above summative equation (2).

It has become clear that constancy in the effect of a fixed change should in no case be expected, unless certain prerequisites are met. One of these clearly is that x_1 (the condition in which the fixed change is introduced), must be a condition variation in which does not entail an accompanying change in any other condition, the magnitude of which accompanying change depends upon the status of the remaining conditions. The following paper by H. Shurrager probably constitutes a good illustration of this difficulty. The effect of a fixed difference in the time between the middle of a series of nonsense syllables and the initiation of recall was kept constant. The effect of this fixed change on recalling syllables in lists of widely different length was then investigated. This procedure, however, resulted actually in the determination of the effect, at different lengths of series, of a fixed change *plus* a variable change. The fixed change was the one already indicated, while the variable change, one which varied with the length of the list, was the time elapsing between the presentation of the *last* syllable of the list and the initiation of recall.

Even where the objective conditions are fully under control, subjective conditions may not be; and constancy in the effect of

a given objective change under various constellations of conditions might well be destroyed by a change in the total constellation of conditions due to change in the subject's attitude, his way of performing the task, or in his motivation. If, when conditions are difficult, the subject performs the task in one way, and when they are easy, in a different way, then there is no reason for expecting constancy in the effect upon *S* of a fixed change in the environmental conditions. For example, at bright illuminations, the subject might try equally hard to see and name each of six widely spaced letters tachistoscopically exposed for .1 sec. At extremely dim illuminations, he might not be aware at all of the outermost letters, and might, therefore, try to see and name only the middle four letters. Now, if the effect of change from 2 to 6 letters at different levels of illumination were constant when the subject, at all levels, tried equally hard for all 6 letters, it would probably not be constant when he tried equally hard for all 6 letters at the higher illuminations but tried only for the middle 4 letters at the lower illuminations.

In view of the above complications, the validity of equations expressing *S* as a sum of functions of its determining conditions, while believed to be indicated in the case of several widely different task-performances by the data so far accumulated, should undoubtedly still be regarded as hypothetical. The problem may, indeed, be regarded largely as one in ways of dividing the total cause of goodness of score into parts, in order to find out whether it is possible to arrive at independently variable part-causes, that is, whether it is possible to formulate, in the case of a given task-performance, a list of determining conditions any one of which may be varied without concomitant variation in any of the remaining determining conditions. The cases which favor an affirmative answer to this hypothesis are mentioned in the latter part of Willmann's paper. His own data, to the effect that the decrease in *S* produced by change from 2 to 4 letters remains unaffected by variation in illumination sufficient to cause the percentage correct scores to vary over most of the possible range, afford a very decisive case in favor of the summation of functions hypothesis. In a number of other cases, however, the effect of

a fixed change was found not to be constant, but to decrease with decrease in the goodness score at which the fixed change was introduced. For example, in a study carried out with the assistance of H. Brogden, it was found that scores in the simultaneous letter-span experiment were less affected by the change from 2 to 4 letters when the exposure-time was extremely short (towards .01 sec.) than when it approached the longer and customary duration of .1 sec. This was the first study in which conditions were used resulting in extreme difficulty (or extremely low goodness of score). This finding led the writer to revise certain previous formulations and to propose a hypothesis to the effect that the change in score produced by a fixed change in conditions depends upon the level of goodness of performance at which it is introduced (3). It was suggested that, while a fixed change in conditions (if not too large) would exert a *nearly* constant effect upon goodness of score as long as the score was high, the effect would progressively decrease, in accordance with a certain hyperbolic curve, as conditions were made very difficult and goodness of score, consequently, became very low. The results obtained by Willmann, however, on the effect of the change from 2 to 4 letters at varying illuminations, seem to prove conclusively that the effect of a fixed change is not dependent upon the level of difficulty at which it is introduced. It is of course possible that whether or not the summative hypothesis holds depends upon what particular conditions are varied. It seems more likely, however, that the summative hypothesis previously formulated is the correct one; and that in those cases in which a fixed change was found not to have a constant effect, the explanation of the inconstancy is to be sought in the occurrence of some additional change, perhaps inadvertent, the magnitude of which varied with the level of difficulty, as in the two illustrations given above. If this view be correct, then it is necessary to assume that in Brogden's study, employing various exposure-times, some uncontrolled condition, either external or internal, varied inadvertently with shortening of the exposure-time and in such a way as progressively to hinder performance in the case of 2 letters more than in the case of 4 letters, or, what amounts to the same thing,

progressively to favor performance in the case of 4 letters more than in the case of 2 letters. Until further investigation, it is impossible to specify with confidence the source of any such uncontrolled variation, though it may be connected with the effect of short duration of exposure upon apparent brightness. That the matter should be reinvestigated, preferably with subjects capable of giving good introspective accounts, is indicated by the contrast between Brogden's data on the effect of the change from 2 to 4 letters with shortened exposure-times and Willmann's data on the effect of this same change with decreased illumination. One might expect that, if, as Brogden found, the change in number of letters has an effect which decreases with shortened exposure-time, such effect would also decrease with lowered illumination. Willmann found, however, rather definite constancy in the effect of the change from 2 to 4 letters at all illuminations except the very brightest.

Further studies in this field are either under way or contemplated. One of these is an attempt to clear up what looks like a discrepancy between the results of Brogden and Willmann. Other problems arise when equation (2) is considered in connection with variations in ability and in amount of practice or learning. Thus, instead of asking what is the effect of a given change in conditions under different constellations of the various objective conditions, it is hoped that it will be possible to determine how the effect of a given change in conditions varies with ability and with stage of learning. While ability cannot be experimentally controlled, different amounts thereof can in most cases be readily obtained by using subjects of different age. Data on the effect of a fixed change in objective conditions of difficulty at various stages of practice have already been obtained and are now being analyzed.

References

1. Woodrow, H. The interrelationship of conditions of difficulty. *J. gen. Psychol.*, 1937, **16**, 83–102, 103–130.
2. ——— Two quantitative laws relating to goodness of performance. *J. of Psychol.*, 1937, **4**, 139–159.
3. ——— The relation between goodness of performance and favorableness of conditions. *Amer. J. Psychol.*, 1938, **51**, 665–677.

THE EFFECT OF A FIXED CHANGE IN DIFFICULTY INTRODUCED AT VARIOUS LEVELS OF DIFFICULTY

By

JOHN M. WILLMANN

The present experiment supplements a series of experiments by Woodrow (2, 3, 4, 5, 6) dealing with the general problem of the relation between goodness of performance and the favorableness of conditions. The problem has been to discover how the goodness of performance on some task is affected by a fixed change in the difficulty of the task, when that change is introduced at various levels of difficulty, these different levels being produced by variations in some other one of the conditions under which the task is performed. Certain discrepancies had appeared in the results of the previous experiments and they were considered to be due perhaps to the fact that the investigations had involved different portions of the total range of difficulty. It was particularly desired then, in the present case, to introduce the fixed change in difficulty at levels of difficulty that would extend all the way from the very easy to the extremely difficult.

The task to be performed consisted in the naming of letters exposed tachistoscopically for .1 second. The letters were projected upon a screen, and the task was rendered more and more difficult, through seven different steps, by the simple process of decreasing the illumination. At each of these levels of difficulty the effect of two different fixed changes was observed. The one change was produced by increasing the separation of two letters, the other by increasing the number of letters from two to four. The problem was to find how these changes in difficulty would affect performance at each intensity of illumination.

The room in which the experiment was conducted was completely darkened so far as external light was concerned, and the subjects were made to sit in total darkness for a period of seven minutes each day before actual work was begun. With the eyes thus quite well dark-adapted, the room was internally

lighted by a small seven-watt bulb, which minimized the effects of whatever further adaptation might occur after the seven minutes.

The subjects, four in number, were employed two at a time, seated at a table with their chins resting in the grooves of two flat horizontal blocks of wood which were clamped to the table, and which were adjustable in height to suit the height of the subjects. With a subject thus seated, his eyes were exactly ten feet from the screen. The projector, a 500-watt Spencer Delineascope, was placed immediately behind the subjects so that the beam passed between their heads. The projector was 33 inches farther from the screen than the subjects' eyes. It was fitted with a special tin holder into which the paper slips containing the letters to be projected were inserted, so that the letters always had the same position when projected upon the screen.

Immediately in front of the lens of the projector was a Whipple tachistoscope, stripped of everything except the disc and the weighted arm to make it revolve. The lens of the projector was 3½ inches in diameter, and the aperture in the tachistoscope was 5½ inches. This aperture was such as to allow an interval of .1 second, as measured by the record of a tuning fork on a smoked paper, from the moment the lens was half uncovered until, after the full exposure of the lens, it was again half covered.

The screen was of white cardboard, 42 inches long and 18 inches high. In the center was a small black cross serving as a fixation point. At the signal of "ready", the subjects looked at this point, and the letters, whether close together or at the wider separation, and whether two or four in number, were projected symmetrically to the right and left of this point. The letters were Willson's gummed letters pasted on white slips of paper, and when projected upon the screen were 1⅛ inches in height. And the spread between the letters of the widest separation when projected was 12¾ inches. Not all the letters of the alphabet were used because of the confusing similarities that exist between certain letters. The omissions were D, F, G, Q, and R. The subjects, however, were not informed of these omissions. The illumination was controlled by two resistors in series which could be set so as to reduce the voltage to any desired point from 110 down to 27.

The subjects were tested on three kinds of material: two letters with a separation on the projection screen of 4¼ inches, two letters with a separation of 12¾ inches, and four letters which actually were a combination of the other two types as shown in the following illustration:

Narrow separation		X	M	
Wide separation	A			P
Four letters	C	O	V	W

The slips on which the letters were pasted were assembled in sets of twenty each, and each set was run off without a break, just a few seconds elapsing between successive presentations, the time required for the experimenter to unload and reload the projector. During these brief intervals the subjects wrote down the letters they thought they saw. They were instructed

to do their best to see the letters correctly but that it was permissible to guess if they had even the faintest idea as to what the letters were. For a correct response it was required not only that a letter be named correctly but that it also be given its correct position in the group.

The seven intensities of illumination mentioned above were selected after two separate hours of preliminary testing, which served incidentally as a practice period. The data obtained from these testings were used to determine what upper and lower limits of illumination would make the task very easy in the one case and extremely difficult in the other, and to locate five intermediate intensities between these extremes which would be more or less evenly spaced as to difficulty. The intensities finally selected were those produced by currents of 50, 44, 37, 35, 33, 31, and 29 volts.

In the experiment itself, then, the subjects were tested at seven different intensities and with three different types of letter groups, making 21 combinations of conditions in all. In each of these combinations each subject was tested to the extent of 200 presentations. Not all intensities and types of letter groups could be tested on any one day, but the 21 combinations were distributed quite uniformly over the entire eighteen days of the experiment. By the end of the fourth day, each combination had been employed to the extent of 40 presentations; by the end of the ninth day, to the extent of 100 presentations; and the second nine days of the experiment duplicated the first nine. Also, the combinations of conditions were mixed or alternated considerably within each day, so that none of them might have an advantageous or disadvantageous position resulting from practice effects or from fatigue. A sufficiently large number of letter slips were available that the subjects could not become familiar with the groups of letters because of frequent repetitions.

The raw scores obtained, consisting of number of letters correctly named, were corrected for chance successes by the use of the formula (1), $S=R-\frac{W}{n-1}$, in which S is the corrected score, R the number of right responses, W the number of wrong

responses, and n the number of alternative responses. Since the subjects did not know that certain letters were never used, n was taken as 26, the number of letters in the alphabet. The corrected scores were next converted into percentage scores.

These percentage scores were then treated in such a way as to render them a more accurate measure of difficulty. In the method used, which has been discussed in detail in another paper (**2**), the simple assumption is made that the amounts of ability brought to bear upon the performance at successive trials by a single subject are normally distributed. On this assumption, it is possible to evaluate the difficulty of the task under any given set of conditions, and thus overcome an imperfection inherent in per cent scores. As here used, difficulty must be conceived to vary inversely with the amount of ability required under the given conditions to produce a correct response. Difficulty has therefore been evaluated on the basis of a normal distribution surface, by using as its measure a distance along the base-line measured in terms of the standard deviation, from the mean to the point at which a perpendicular would mark off a proportion of the surface equal to the proportion of correct responses. Thus, if under one set of conditions a subject correctly named 16 per cent of the letters and under another set of conditions 50 per cent, the difficulty for that subject of naming a letter correctly would be regarded as $x+0$ in the latter case and as $x+1.0$ in the former case. Similarly, the difficulty of any set of conditions may be evaluated in an absolute way from an arbitrary zero, provided only the score lies within the limits of zero and one hundred per cent correct.

Since the variability in the ability of different subjects would presumably not be the same, the data for each subject were in the first instance treated separately. If, however, it is a valid procedure to transform the per cent correct scores of a single subject into sigma values, it is probably sound, also, to transform the average per cents correct of all subjects used, since such average per cent correct scores are likely to be somewhat more representative of the scores made by any individual belonging to the same population from which the used subjects were chosen. An alter-

native procedure, which has a nearly identical outcome, is to average the sigma scores. Since the complete experiment, gone through by only four subjects, involved twenty-one sets of conditions, the results presented in this case will be limited to the average per cent correct scores and their sigma values. In the case of a supplementary experiment carried out with twelve subjects but involving only four sets of conditions, both individual and pooled scores will be presented.

The results of the complete experiment with four subjects are given in Table I. This table required the scoring of a total of 44,800 letters, of which one half, or 22,400, were exposed two at a time and the other half four at a time. The percentage correct in each of the 21 cells of the table is based on 800 exposures, and therefore is a percentage calculated from 1600 in

TABLE I

PER CENT CORRECT AND SIGMA SCORES UNDER TWENTY-ONE CONDITIONS

	Two Letters Narrow Separation		Two Letters Wide Separation		Four Letters	
Illumination	%	σ	%	σ	%	σ
I	90.90	—1.34	78.16	—.78	72.80	—.61
II	83.36	— .97	58.08	—.20	62.43	—.32
III	49.82	.01	29.67	.53	30.78	.50
IV	40.33	.25	20.57	.82	22.46	.76
V	25.64	.65	9.13	1.33	11.83	1.18
VI	10.76	1.24	3.80	1.77	4.39	1.71
VII	3.87	1.77	.68	2.47	1.33	2.22

the case of two-letter exposures and from 3200 in the case of four-letter exposures. Each per cent correct given in the table is the per cent correct for the pooled responses of the four subjects after the total number correct has been corrected for chance successes by the formula already given.

The basic problem is the drop in performance level, resulting at each illumination, from the change from the narrow separation to the wide, and from two letters to four. The change from two letters to four letters requires special treatment. We are concerned with the effect of a *single* change in the conditions when introduced at various levels of difficulty that are produced by some other change in conditions. Now the shift from two letters

to four is not a single change, for in addition to the change in number of letters there is also an alteration in the spread or position of the letters. No two letters can duplicate the pattern of four letters. It is the effect of the change in number alone that is sought, and the disturbing spatial factor must be eliminated. This was done by averaging the percentages correct for both narrow and wide 2-letter separations before taking the sigma value to be compared with that of the 4-letter series at the same intensity of illumination. For the 4-letter group, as already mentioned, is really a combination of the two letters at the narrow and at the wide separations. It follows that the change from the average of both 2-letter series to the 4-letter series represents

TABLE II

CHANGE IN DIFFICULTY AT VARIOUS ILLUMINATIONS DUE TO A FIXED CHANGE IN EITHER SEPARATION OR NUMBER OF LETTERS

Illumination	Change in Separation (from narrow to wide)	Change in Number (from two to four)
I	.56	.41
II	.77	.23
III	.52	.24
IV	.57	.25
V	.68	.25
VI	.53	.25
VII	.70	.22

that change from " 2-ness " to " 4-ness " which is desired. Similarly, the fixed change from two letters at the narrow separation to two letters at the wide separation represents only a change in separation, without the factor of number of letters entering in.

The effect of these two fixed changes is shown in Table II. The first column of differences represents the increase in difficulty (sigma score), at each intensity, caused by increasing the separation of the two letters, and the second column the increase in difficulty caused by the change from two letters to four letters.

In the shift from the narrow to the wide separation the drop in performance level, as measured by the sigma values, is irregular; but when viewed across the whole range of intensities, the data indicate constancy of effect rather than any general trend towards either increase or decrease. The best fitting straight line would be practically horizontal. It is concluded, therefore, that

the data obtained indicate a general tendency towards constancy in the effect of a fixed increase in the separation of two letters, no matter at what illumination the fixed increase is introduced.

The effect of the change from 2 to 4 letters is slightly less certain. The results indicate a rather marked decrease in the effect of the increase in number of letters as the illumination drops from level I to level II, but a remarkable constancy in the effect of this fixed change in number throughout levels II to VII inclusive.

TABLE III

PERCENTAGE AND SIGMA SCORES FOR TWO LETTERS UNDER FOUR SETS OF CONDITIONS

Illumination	Bright				Dim			
Separation	Narrow		Wide		Narrow		Wide	
Subject	%	σ	%	σ	%	σ	%	σ
F	87.98	—1.17	69.13	—.50	51.58	—.04	28.83	.56
K	91.88	—1.40	78.23	—.78	62.30	—.31	30.78	.50
De	78.23	— .77	52.88	—.07	26.23	.64	15.50	1.01
Dr	71.08	— .55	42.23	—.05	22.00	.77	10.95	1.23
J	87.00	—1.13	65.88	—.41	42.15	.20	21.35	.79
B	89.28	—1.24	71.73	—.57	32.73	.45	15.83	1.00
H	37.60	.32	23.63	.72	3.80	1.77	2.18	2.01
Mo	91.88	—1.40	78.88	—.80	51.25	—.03	26.55	.63
Mc	89.28	—1.24	77.25	—.75	60.68	—.27	26.55	.63
W	83.43	— .97	65.88	—.41	37.28	.32	14.85	1.04
O	88.30	—1.19	66.53	—.43	36.95	.33	18.45	.90
S	83.75	— .97	62.63	—.32	28.50	.57	5.75	1.58
Pooled Scores	81.74	— .906	63.73	—.351	37.95	+.307	18.13	+.910
Mean of σ Scores		— .976		—.364		+.367		+.990

The first of the above conclusions, to the effect that a given increase in separation tends to exert a constant effect at all illuminations, for reasons to be explained later is considered to be extremely important. It was therefore subjected to further test by a supplementary experiment employing twelve subjects. This experiment differed from the preceding one in that only two intensities of illumination were used, namely, II and IV of the preceding experiment, and in that only 2-letter slips were used. Both the average scores made by the group and the individual scores are shown in Table III. One hundred and sixty slips were presented to each subject under each of the four sets of conditions. It follows that each individual score is the percentage

correct (corrected by formula) of 320 letters. The average percentage correct (also corrected) given in the second line from the bottom is in each case the percentage correct of 12×320 letters or 3,840 letters. The entire table required the scoring of 15,360 letters.

It will be observed from Table III that the difference between illuminations II and IV was sufficient to produce a great difference in difficulty. Thus, with the brighter illumination the

TABLE IV

THE EFFECT OF A FIXED CHANGE IN SEPARATION OF LETTERS AT TWO DIFFERENT LEVELS OF ILLUMINATION

Subject	Bright (II)	Dim (IV)	Diff. (IV–II)
F	.67	.60	—.07
K	.62	.81	.19
De	.70	.37	—.33
Dr	.50	.46	—.04
J	.72	.59	—.13
B	.67	.55	—.12
H	.40	.24	—.16
Mo	.60	.66	.06
Mc	.49	.90	.41
W	.56	.72	.16
O	.76	.57	—.19
S	.65	1.01	.36
Pooled Scores	.555	.603	.048
Mean of differences in sigma scores			.012
$\sigma_{diff.}$ of differences			.218

average percentage correct for the narrow separation was 81.74 while with the dimmer illumination it was only 37.95. The effect of this change in level of difficulty upon the effect of the fixed change in separation of the letters is shown by Table IV. The first column of Table IV, headed "Bright (II)", shows the increase in difficulty (in sigma scores) produced by the increased separation when the illumination was relatively bright (at level II); and the second column, the same effect when the illumination was relatively dim (at level IV). The third column shows the difference in the effect of increased separation at the two levels of illumination. This column reveals that individual differences are rather pronounced. With some subjects the effect of the increase in separation is greater with bright illumination and with

others at dim illumination. On the average, however, there is no significant difference, since the mean discrepancy between the effect of increased separation at the two illuminations (.012) is only about one-nineteenth of the standard deviation of the distribution of the 12 cases (.218).

It happens that, in the present instance, even in terms of percentage correct scores (pooled scores), the effect of increase in separation may be seen to be practically the same at both levels of illumination. The reason that this is possible is presumed to lie in the fact that the lowest percentage score (18.13, with illumination IV and wide separation) is above zero almost the same amount that the highest percentage score (81.74, with illumination II and narrow separation) falls short of one hundred. It thus happens that the changes in the percentage scores are comparable. Change from narrow to wide separation with illumination II causes a drop in the percentage correct scores from 81.74 to 63.73, or 18.01, while the change from wide to narrow separation at illumination IV causes an increase of nearly equal magnitude—one from 18.13 to 37.95, or 19.82.

It may be concluded, then, that the supplementary experiment with twelve subjects verifies the conclusion drawn from the more elaborate investigation with four subjects as regards the effect of a fixed change in letter separation at different levels of difficulty when the latter are produced by variation in illumination. Both experiments indicate that the effect of the fixed change tends to remain constant at the different levels of difficulty.

As already stated, the problem here studied is identical with one studied in a series of investigations by Woodrow, namely, that of the effect of a fixed change in conditions at various levels of difficulty. Certain inconsistencies had appeared upon which it was hoped some light would be thrown by the present experiment. One of these previous experiments (4) dealt with reaction-time. In that experiment the subject's task consisted of reacting as quickly as possible in response to a flash of light. Different levels of difficulty were produced by varying the intensity of this light throughout a wide range. The fixed change introduced at these different levels of difficulty consisted in the shift from a

standard preparatory interval (the time between the "ready" signal and the flashing of the light) of exactly 2 sec. duration, to a variable preparatory period ranging from 2 to 24 sec. The reaction-time scores in thousandths of a second were changed, by means of an absolute scaling technique, into scores representing goodness of performance. It was then found that the effect of the fixed change in the length of the preparatory period exerted a constant effect upon goodness of performance at all the different levels of difficulty produced by varying the intensity of the light. This result is in harmony with the findings in the present experiment.

In another investigation (3:I) the procedure was quite similar to that followed in the present experiment. The subjects were required to name letters exposed tachistoscopically. The fixed change in conditions was a shift from two letters to four, but the different levels of difficulty at which this change was introduced were produced, not by decreasing illumination, but by increasing the distances between the letters. Here again constancy was found for the effect of the fixed change from two to four letters introduced at the various levels of difficulty.

In a subsequent experiment (3:II), however, some different results were obtained, partly agreeing and partly disagreeing with the previous findings. In this experiment changes in three conditions were utilized, namely, the number of letters, the illumination, and the spacing between the letters. In this study, the effect of a fixed change did not always remain constant, but sometimes decreased with increase in difficulty. The author, however, concluded as follows: "It follows that in any experimental work in which the investigator finds that a fixed change does not exert a constant absolute effect at all levels of difficulty, he should immediately search for a factor which he has not kept constant . . . which varies . . . with change in the conditions variation in which produced the different levels of difficulty at which this fixed change was introduced" (3:II, p. 128).

The general conclusions just stated were later modified because of results obtained in still another experiment (6) in which a new change in conditions was used, namely, the length of time the

letters were exposed. Exposures of .100, .020, .015, and .010 sec. were used. These variations produced the different levels of difficulty, and the fixed change introduced at each level was the change from 2 to 4 letters. The fixed change had a decreasing effect as the task became more difficult. This outcome suggested a new explanation of the fact that under some conditions increase in difficulty lessened the effect of a fixed change in conditions. This explanatory hypothesis rested upon the fact that a review of the previously obtained data revealed that constancy in the effect of a fixed change in conditions had been obtained only at relatively high levels of performance and with relatively small fixed changes. The hypothesis was therefore advanced that while the effect of a fixed change is constant, or nearly so, at easier levels, it becomes progressively smaller with further increases in difficulty once a high level of difficulty is reached. Such results would be accounted for, it was pointed out, on the basis of a certain curvilinear relationship between scaled scores and the favorableness of conditions.

The present experiment was undertaken with a view to testing this latter hypothesis by investigating levels of difficulty ranging from the very easy to the extremely difficult. These latter levels were certainly reached, for at the dimmest illumination scores little better than pure chance were secured. Yet throughout this wide range in levels of difficulty, the effect upon difficulty of the fixed increase in separation of two letters showed no general tendency to decrease with increase in difficulty. It follows that the relationship between goodness of performance and favorableness of conditions is linear, or very nearly so. It seems, therefore, that those cases in which a decrease occurs in the effect of a fixed change with increase in difficulty can be explained only in terms of Woodrow's original hypothesis concerning them, briefly indicated by the citation given above.

The results here obtained concerning the fixed change from 2 to 4 letters, with only slightly less certainty, indicate the same conclusion. There did occur, however, a decrease between the highest and next highest illumination in the effect of this change. Moreover, it is quite possible that had the change been from 2 to

6 letters there would have been observed a more marked decrease in the effect of the change in number with decrease in illumination. Such a result would be in line with Woodrow's finding that the change from 2 to 4 letters had a constant effect at various levels of difficulty produced by variation in either the separation or the illumination of the letters, but that the change from 2 to 6 letters had an effect which decreased with increase in the level of difficulty.

References

1. GUILFORD, J. P. The determination of item difficulty when chance success is a factor. *Psychometrika,* 1936, **1**, p. 260.
2. WOODROW, H. The measurement of difficulty. *Psychol. Rev.,* 1936, **43**, 341–365.
3. ——— The interrelationship of conditions of difficulty: I. The effect of change in number at various spatial separations on simultaneous letter span. *J. gen. Psychol.,* 1937, **16**, 83–102; and II. Number, spatial separation, and illumination as conditions of simultaneous letter span. *Ibid.,* 1937, **16**, 103–130.
4. ——— Two quantitative laws relating to goodness of performance. *J. Psychol.,* 1937, **4**, 139–159.
5. ——— The effect of pattern upon simultaneous letter span. *Amer. J. Psychol.,* 1938, **51**, 83–97.
6. ——— The relation between goodness of performance and favorableness of conditions. *Amer. J. Psychol.,* 1938, **51**, 665–677.

THE MEASUREMENT OF MEMORY ON AN ABSOLUTE SCALE

By

HARRIETT C. SHURRAGER

Problem.—Among the conditions which affect the difficulty of recalling a nonsense syllable are (1) the length of the series in which the syllable occurs and (2) the time interval intervening between presentation and recall. In the present experiment the difficulty (on the average) of naming any syllable in a list was varied by means of changes in the above two conditions. Length of list was varied from 3 to 36 syllables and the time of recall was either 116 sec. (measured from the presentation of the middle syllable of a list) or two days, both intervals being used with each of the ten different lengths of list. In all, therefore, the recall of the syllables was measured under twenty different sets of conditions.

The chief purpose of the study was to determine whether the difficulty due to these various conditions could be measured in units which would permit comparison of the difference in difficulty under any pair of sets of conditions. Such comparisons require the measurement of difficulty in constant or equal units, and are impossible, as is universally recognized, as long as the data indicate only the mean per cent correct score under each set of conditions. To determine the difference in difficulty between any two sets of conditions so as to compare the difference with that obtaining under any other two sets of conditions, it is necessary to measure the difficulty in terms of constant units. Such units can be obtained only by the technique known as scaling. This technique was applied to the measurement of difficulty by Woodrow,[1] who used Thurstone's scaling technique for that purpose. The first problem envisaged by the present study, then, was whether the Thurstone-Woodrow procedure was applicable to memory, when the raw scores consist of the percentages of nonsense syllables recalled under various conditions.

Whether absolute scaling is possible depends upon whether a

[1] For Thurstone's method of absolute scaling see (3) and for Woodrow's application of this method to the scaling of difficulty, see (4).

linear relationship is obtained when the percentages of the population exceeding all various possible scores made under one set of conditions is plotted against the percentages exceeding the same scores under a second set of conditions, after all percentages have been transformed into x-values. By an x-value is meant the distance along the base-line of a normal distribution, measured from the mean, to the point at which a perpendicular would mark off an area to the right of the line equal to the percentage of the population exceeding a given score. Since such relationships were found to be linear, it was concluded that the difficulty of remembering a nonsense syllable under different sets of conditions could be determined by absolute scaling. In view of this finding, three experimental problems were considered.

(1) The relation of scaled difficulty to length of list.

(2) The effect of a fixed change in difficulty at different levels of difficulty. The fixed change employed was a delay of two days in recall. The question asked, then, was whether a delay of two days between presentation and recall increased the mean difficulty per recalled syllable, in the case of short lists by the same amount as in that of long lists.

(3) The relation between difficulty and variability.

Procedure.—The subjects were 766 university students enrolled in twenty-seven sections of Psychology I. These twenty-seven sections were divided into three groups of nine sections each. Group A was tested on lists of 3, 6, and 9 syllables in the order 3-3-9-6-6-6-9-3-3; group B on lists of 12, 15, and 18 syllables in the order 12-15-18-18-15-12; and group C on 21, 24, 30, and 36 syllables in the order 21-24-30-36-30-24-21. Since it was deemed advisable, in order to avoid interference of lists, to show the subjects only one list of syllables a week, it was necessary to use three groups of subjects in order to give all of the various lists in the limited time of a single semester. The shorter lists were presented more often than the longer ones so that, as far as feasible, by reducing the variation in the total number of syllables from which the percentages correct for the various lengths of list were calculated, the reliability of these percentages would not depend in too extreme a manner upon the length of the list. A tentative assumption of the study is that the three groups of subjects were

equal in ability. The comparative smoothness of the obtained curves (Fig. 16) representing increase in absolute difficulty with increase in length of list seems to indicate the approximate correctness of this assumption of equal ability in the three groups. In so far as the assumption is not correct, the apparent difference in difficulty between two lengths of list may be due to difference in ability of the groups.

The nonsense syllables used were chosen from Glaze's list (1) of 3-letter syllables. Each syllable was accurately pasted on a 10-inch white cardboard square with 2-inch Wilson, black, gummed letters. The lists were so arranged that the same vowel or consonant did not occur in two successive syllables. A blank card was inserted between each two syllables and each card exposed for two seconds. The rate of exposure was kept constant with the aid of a metronome.

At the beginning of the class hour (on Monday or Tuesday) the experimenter showed the students a list of syllables. They were told that they would be asked to write down the syllables they recalled after the entire list had been shown and again at the next class meeting (two days later).

A student's score for a list of a given length was included in the data, provided he was present each time that length of list was shown and regardless of his absences when other lengths of list were presented. It follows that no two distributions of scores represent exactly the same group of subjects or the same number of subjects. Absences from class are so frequent that it is virtually impossible to make 100 per cent attendance the criterion for including a student's score in the data. Comparison of the averages obtained from 35 students in Group A who were never absent with those obtained from the larger groups which included students who were occasionally absent showed that the differences were small and irregular. On the basis of this result, and in view of the greater reliability of percentages calculated from a large group, the procedure of including a student's score in the data provided he was present every time that length of list was shown, though not strictly justified, was the one followed.

In calculating the number of individuals exceeding any score, it was assumed that half the number making that score really

exceeded it and that half failed to exceed it. All scores of 100 per cent correct are tabulated simply as exceeding 97.5 per cent.

The method here used to determine the mean difficulty of each of the twenty sets of conditions is what Woodrow (4) has called the "population percentage" scaling method. So far as the mathematical procedure is concerned it is identical with Thurstone's scaling method. The difference is one of application and interpretation. Thurstone used the method in the case of different groups tested by the same tests. As thus employed, it yields the mean difference in scaled amount of ability between the two groups. In the present study, the method is applied to the case of a given group tested under different determining conditions of scores, and the results are interpreted as indicating the change in difficulty due to change in these conditions under which the task is performed.

Thurstone (3) has shown that, in the case of two distributions, for sets of x-values between which the correlation is +1.0, the following equation may be written:

$$\sigma_b = \frac{S_a}{S_b}\sigma_a \qquad (1)$$

in which S_a and S_b represent standard deviations of the x-values and σ_a and σ_b absolute variability in terms of the standard deviation of the easiest set of conditions. He has also shown that the difference between the means of two overlapping distributions may be expressed as follows:

$$M_b = \sigma_a \left(m_a - \frac{S_a}{S_b} m_b \right) + M_a \qquad (2)$$

where M_a and M_b represent means and m_a and m_b are means of the x-values of a given set of scores occurring in both distributions.

Applied to data obtained from two groups of subjects tested under identical conditions, equation (2) gives the difference in mean ability between the two groups. Applied to data obtained from a single group of subjects tested under different conditions, (2) gives the difference in mean difficulty of the two sets of conditions. By mean difficulty, in the present instance, is meant the mean for all the subjects of the difficulty of naming a nonsense syllable under any one set of conditions. By the difficulty of

naming a nonsense syllable is meant, of course, not the difficulty in the case of any particular syllable, but the mean difficulty for all the syllables used.

Results.—The data are given in sufficient detail in Tables I and II, which shows the x-values of the percentages of the population exceeding any of the obtainable scores listed in the left-hand column, headed " Per Cent Correct ".

TABLE I

IMMEDIATE RECALL: STANDARD DEVIATION VALUES OF THE PERCENTAGES OF THE GROUP EXCEEDING THE GIVEN PER CENT CORRECT SCORES

Per Cent Correct	Number of Syllables				
	3	6	9	12	15
97.5	—1.185				
95.0	—1.019				
92.5	— .827		—2.748		
90.0	— .631		—2.652		
87.5	— .470	—2.652	—2.576		
85.0	— .319	—2.457	—2.576		
82.5	— .179	—2.290	—2.512		
80.0	— .040	—2.097	—2.366		
77.5	+ .098	—1.960	—2.226		
75.0	+ .238	—1.581	—2.075		
72.5	+ .375	—1.483	—1.977		
70.0	+ .519	—1.353	—1.911		
67.5	+ .671	—1.185	—1.866		
65.0	+ .789	—1.011	—1.812		
62.5	+ .881	— .845	—1.751		
60.0	+ .986	— .678	—1.665		
57.5	+1.094	— .527	—1.555		
55.0	+1.203	— .377	—1.439		
52.5	+1.317	— .235	—1.276	—2.257	
50.0	+1.461	— .060	—1.141	—2.033	
47.5	+1.499	+ .113	— .974	—1.787	—2.409
45.0	+1.762	+ .274	— .831	—1.589	—2.120
42.5	+2.014	+ .440	— .625	—1.412	—1.911
40.0	+2.120	+ .595	— .421	—1.190	—1.706
37.5	+2.366	+ .749	— .230	— .990	—1.499
35.0	+2.748	+ .923	— .013	— .772	—1.259
32.5		+1.067	+ .184	— .568	—1.007
30.0		+1.248	+ .372	— .356	— .749
27.5		+1.514	+ .568	— .138	— .513
25.0		+1.675	+ .755	+ .085	— .279
22.5		+1.881	+ .994	+ .311	— .048
20.0		+2.197	+1.136	+ .547	+ .253
17.5		+2.409	+1.335	+ .820	+ .434
15.0		+2.878	+1.522	+1.076	+ .722
12.5			+1.762	+1.359	+1.041
10.0			+2.054	+1.555	+1.353
7.5			+2.326	+1.787	+1.695
5.0			+2.878	+2.075	+2.170
2.5			+3.090	+2.366	
0.0				+2.748	

TABLE I—*Continued*

Per Cent Correct	Number of Syllables				
	18	21	24	30	36
42.5	—2.409				
40.0	—2.120				
37.5	—1.896				
35.0	—1.866	—2.457	—2.120		
32.5	—1.412	—2.054	—2.033		
30.0	—1.185	—1.774	—1.717	—2.033	—2.457
27.5	— .978	—1.476	—1.499	—1.685	—1.866
25.0	— .782	—1.117	—1.243	—1.311	—1.433
22.5	— .539	— .762	— .893	— .999	—1.063
20.0	— .243	— .457	— .519	— .681	— .745
17.5	+ .058	— .113	— .192	— .277	— .440
15.0	+ .385	+ .233	+ .121	+ .136	— .092
12.5	+ .759	+ .539	+ .479	+ .601	+ .269
10.0	+1.160	+ .911	+ .813	+ .982	+ .631
7.5	+1.546	+1.323	+1.126	+1.185	+1.019
5.0	+1.997	+1.774	+1.522	+1.762	+1.655
2.5	+2.576	+2.366	+1.911	+2.512	+2.097
0.0		+2.457	+2.409		

TABLE II

DELAYED RECALL: STANDARD DEVIATION VALUES OF THE PERCENTAGES OF THE GROUP EXCEEDING THE GIVEN PER CENT CORRECT SCORES

Per Cent Correct	Number of Syllables				
	3	6	9	12	15
97.5	—2.170				
95.0	—2.075				
92.5	—1.881				
90.0	—1.762				
87.5	—1.665				
85.0	—1.572				
82.5	—1.454				
80.0	—1.299				
77.5	—1.170	—2.576			
75.0	—1.058	—2.366			
72.5	— .900	—2.226			
70.0	— .762	—2.097	—2.878		
67.5	— .640	—2.014	—2.748		
65.0	— .553	—1.927	—2.576		
62.5	— .482	—1.825	—2.409		
60.0	— .415	—1.728	—2.290		
57.5	— .345	—1.626	—2.197		
55.0	— .269	—1.538	—2.120		
52.5	— .194	—1.476	—2.014		
50.0	— .121	—1.412	—1.943		
47.5	— .013	—1.335	—1.852		
45.0	+ .098	—1.259	—1.762		
42.5	+ .209	—1.180	—1.675		
40.0	+ .295	—1.098	—1.580		
37.5	+ .372	—1.011	—1.499		

TABLE II—*Continued*

Per Cent Correct	Number of Syllables				
	3	6	9	12	15
35.0	+ .451	— .908	—1.425		
32.5	+ .533	— .810	—1.347	—2.576	
30.0	+ .610	— .703	—1.243	—2.326	
27.5	+ .693	— .604	—1.141	—2.097	
25.0	+ .779	— .496	— .970	—1.927	—2.257
22.5	+ .881	— .391	— .806	—1.461	—1.838
20.0	+ .994	— .253	— .625	—1.126	—1.499
17.5	+1.122	— .115	— .391	— .782	—1.211
15.0	+1.237	+ .048	— .261	— .490	— .919
12.5	+1.347	+ .245	— .083	— .202	— .625
10.0	+1.483	+ .404	+ .121	+ .058	— .327
7.5	+1.616	+ .583	+ .358	+ .353	— .108
5.0	+1.751	+ .681	+ .622	+ .706	— .559
2.5	+1.911	+1.036	+ .938	+1.080	—1.071
0.0	+2.144	+1.385	+1.392	+1.635	—1.762

Per Cent Correct	Number of Syllables				
	18	21	24	30	36
22.5	—2.326		—2.144		
20.0	—2.075	—2.366	—1.911		
17.5	—1.799	—2.075	—1.695	—1.695	—2.226
15.0	—1.499	—1.717	—1.546	—1.372	—1.787
12.5	—1.131	—1.287	—1.276	— .958	—1.405
10.0	— .759	— .786	— .796	— .662	— .990
7.5	— .372	— .259	— .327	— .204	— .521
5.0	+ .110	+ .269	+ .171	+ .388	— .000
2.5	+ .775	+ .885	+ .703	+1.398	+ .595
0.0	+1.655	+1.751	+1.454	+2.366	+1.385

Before proceeding further with the determination of the difference in difficulty due to the change in conditions, it is desirable to examine the relationship existing between the x-values, or difficulty values, of the various scores made under different conditions. The validity of the absolute scaling method depends largely upon the linearity of the relation between the x-values for any two distributions which are plotted against each other. If the x-values beyond the range of $\pm 1.7\sigma$ be excluded, they are shown by Figures 1 to 14 inclusive to give plots closely approximating linearity. The range $\pm 1.7\sigma$ includes all scores not exceeded by more than 95.5 nor less than 4.5 per cent of the population. It is undesirable to use the x-values for percentages lying near the extreme ends of the distributions because of the relatively great unreliability of extreme portions of the distribution surface.

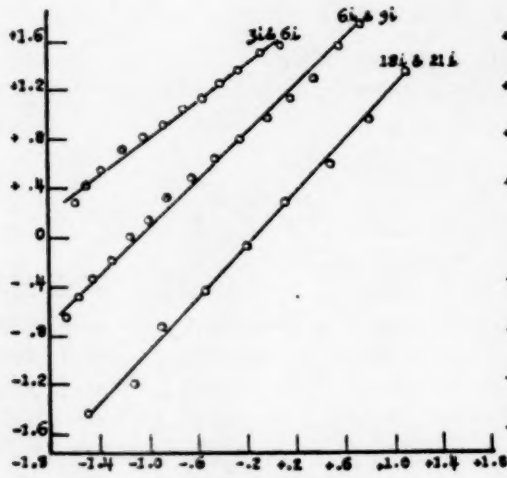

FIG. 1. The relation between x-values of a fixed set of scores under conditions $3i$ and $6i$, and $6i$ and $9i$, and $18i$ and $21i$.

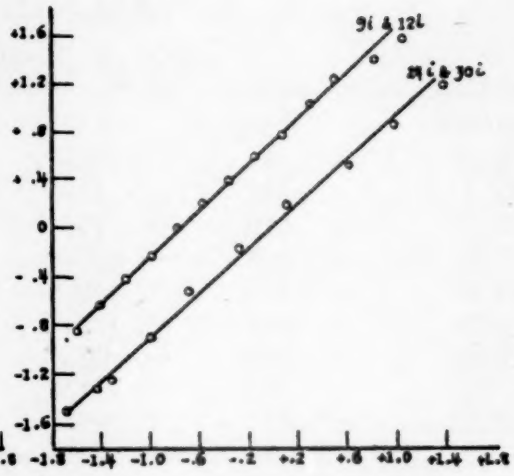

FIG. 2. The relation between x-values of a fixed set of scores under conditions $9i$ and $12i$ and $24i$ and $30i$.

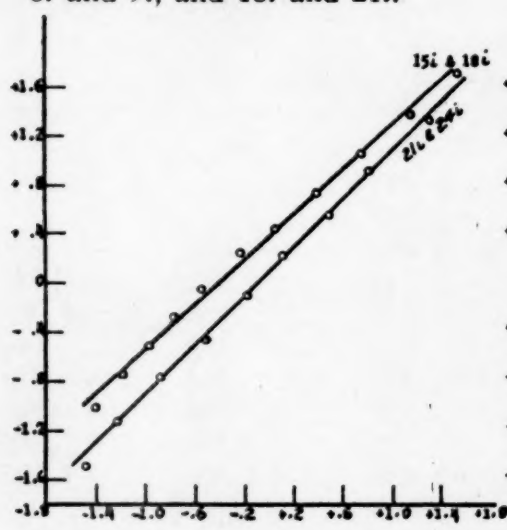

FIG. 3. The relation between x-values of a fixed set of scores under conditions $15i$ and $18i$ and $21i$ and $24i$.

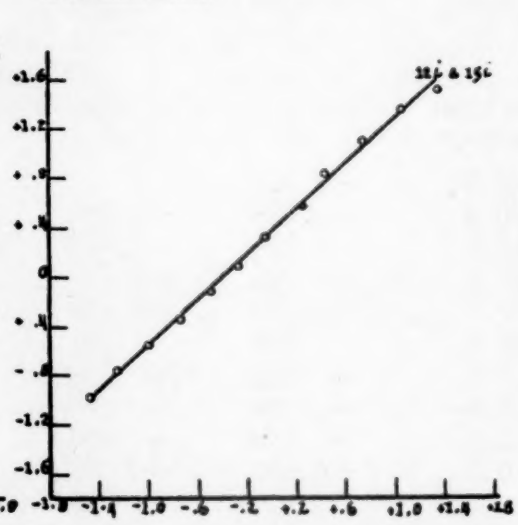

FIG. 4. The relation x-values of a fixed set of scores under conditions $12i$ and $15i$.

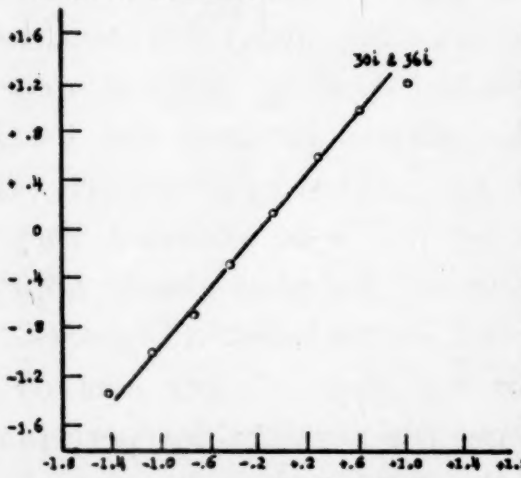

FIG. 5. The relation between x-values of a fixed set of scores under conditions $30i$ and $36i$.

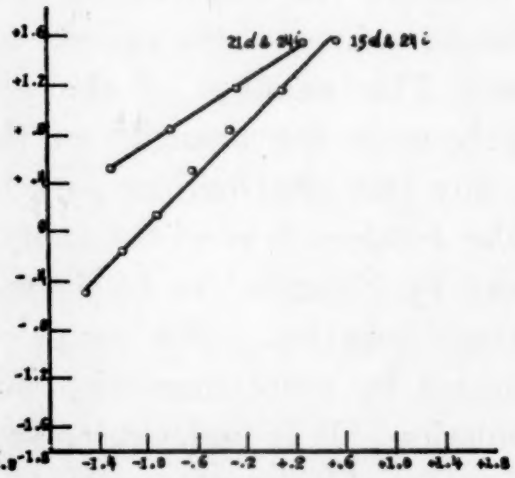

FIG. 6. The relation between the x-values of a fixed set of scores under conditions $21d$ and $24i$ and $15d$ and $24i$.

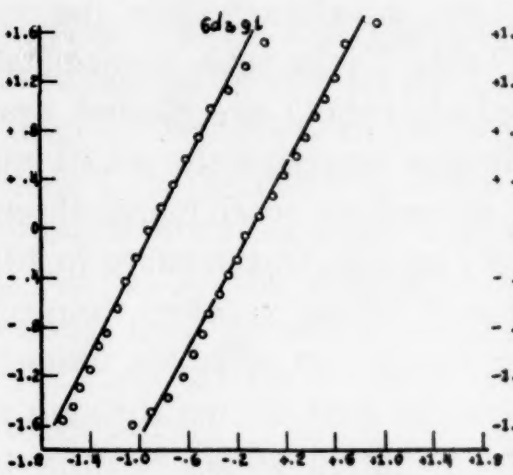

FIG. 7. The relation between x-values of a fixed set of scores under conditions $6d$ and $9i$ and $3d$ and $6i$.

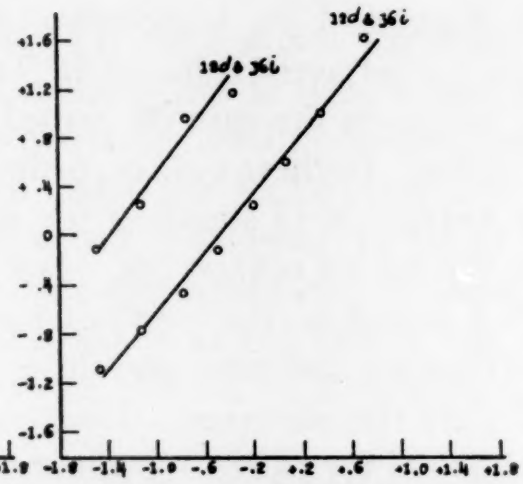

FIG. 8. The relation between x-values of a fixed set of scores under conditions $18d$ and $36i$ and $12d$ and $36i$.

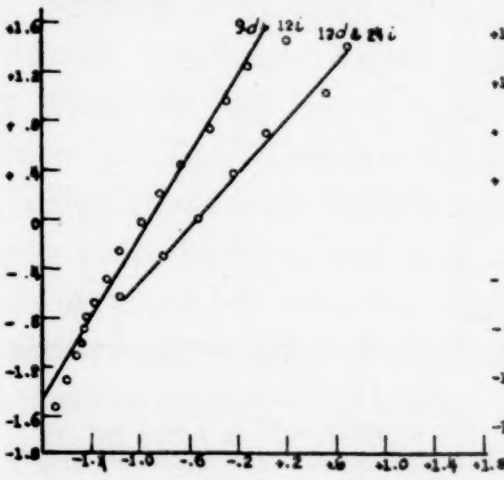

FIG. 9. The relation between x-values of a fixed set of scores under conditions $9d$ and $12i$ and $12d$ and $24i$.

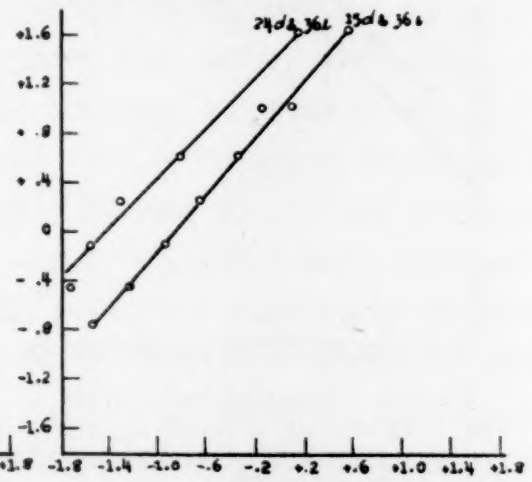

FIG. 10. The relation between x-values of a fixed set of scores under conditions $24d$ and $36i$ and $15d$ and $36i$.

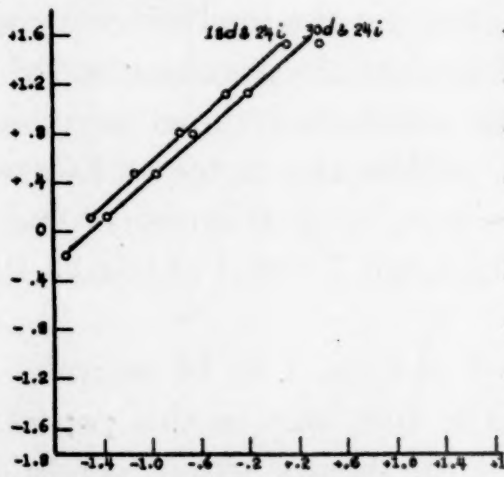

FIG. 11. The relation between the x-values of a fixed set of scores under conditions $18d$ and $24i$ and $30d$ and $24i$.

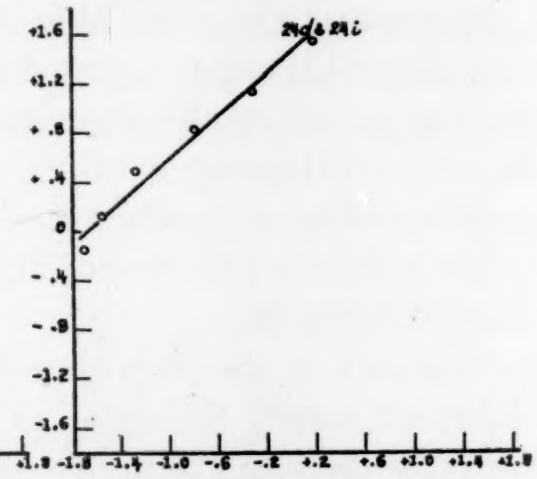

FIG. 12. The relation between the x-values of a fixed set of scores under conditions $24d$ and $24i$.

In Figures 1 to 5 inclusive, the x-values within the range $\pm 1.7\sigma$ for adjacent columns in Table I (the table showing difficulty scores in the case of immediate recall) are plotted against each other. In these figures, ordinates represent the set of conditions nearest $3i$ (i standing for immediate recall) and abscissae those for the set nearest $36i$. For example, in the curve in Fig. 1 labelled $3i$ and $6i$, the x-values for $3i$ (three syllables, immediate recall) are the ordinates and those for $6i$ (six syllables, immediate recall) are the abscissae. The straight lines drawn through each set of points for paired coördinates are fitted simply by inspection.

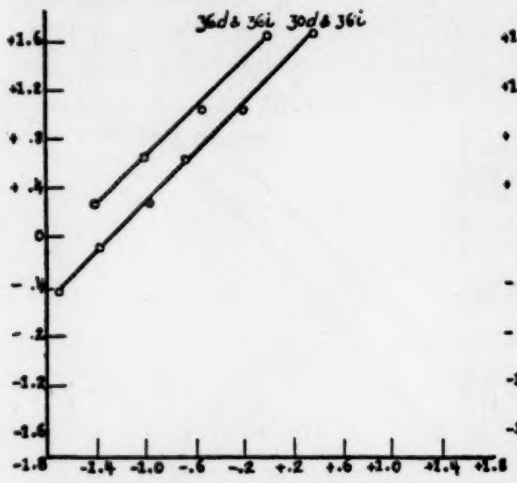

FIG. 13. The relation between the x-values of a fixed set of scores under conditions $36d$ and $36i$ and $30d$ and $36i$.

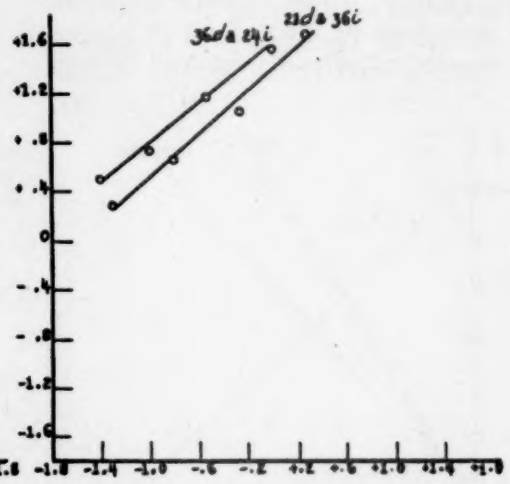

FIG. 14. The relation between the x-values of a fixed set of scores under conditions $36d$ and $24i$ and $21d$ and $36i$.

Figures 6 to 14 inclusive show the relation between x-values under the conditions of immediate and delayed recall. In this case, the x-values given in Table II under the ten different conditions of delayed recall are plotted against the x-values in Table I for the immediate recall lists with which they show the greatest number of overlapping x-values within the range $\pm 1.7\sigma$. In these plots, ordinates represent x-values for distributions of scores under the conditions of immediate recall *(i)* and abscissae those for delayed recall *(d)*.

The linearity of the plots shown in Figs. 1 to 14 indicates the feasibility of scaling the data. The first step in this process is to determine the mean difficulty of the twenty sets of conditions. This is done by applying to each two sets of x-values which are plotted against each other equations (1) and (2), which have

been given above. By equation (1), the absolute variability of the x-values under one set of conditions can be determined when the absolute variability of the other is assumed or known, and by equation (2), the absolute mean under one set of conditions can be obtained when the other is assumed or known. In the present case, the σ_{dis} of the x-values under the conditions of three syllables lists and immediate recall is taken as the unit of the scale and the mean of the x-values under this set of conditions as an arbitrary zero.

In determining the differences in difficulty between conditions under immediate recall, in all cases adjacent distributions were scaled against each other. In determining the differences in difficulty between immediate and delayed recall, the set of x-values in delayed recall was scaled against the set in immediate recall with which it had the greatest number of overlapping points, between $\pm 1.7\sigma$. The last seven lists of delayed recall, 12*d* to 36*d* inclusive, overlap equally with 24*i* and 36*i*. They were scaled against both 24*i* and 36*i;* and since the differences between the values obtained by the two scaling processes were slight, the results as regards both means and standard deviations were averaged.

After the scaling was completed, an examination was made of the relationship between the mean per cent correct scores for the various lengths of list and the scaled mean difficulty values. The graph obtained (Fig. 15) revealed at once that the mean per cent scores varied as the ogive of a normal distribution curve when scaled values were placed along the base-line. Representing the mean percentage correct by p, and the scaled value of this mean score by x, the relation between the two sorts of scores is expressed by the formula.

$$p = \int_{-\infty}^{\frac{x-a}{b\sqrt{2}}} \frac{e^{-\frac{(x-a)^2}{2b^2}}}{b\sqrt{2\pi}}\, dx$$

Rough calculations indicate a value of approximately 1.4 for a and one of approximately 2.0 for b.

Table III gives the absolute means and standard deviations, together with the mean percentage correct, for all twenty constellations of conditions. Fig. 17 is a graphic representation of the relation between mean scaled difficulty and length of list both in the case of immediate recall and that of recall after two days. All values represented in Figures 16 and 17 are given in Table III.

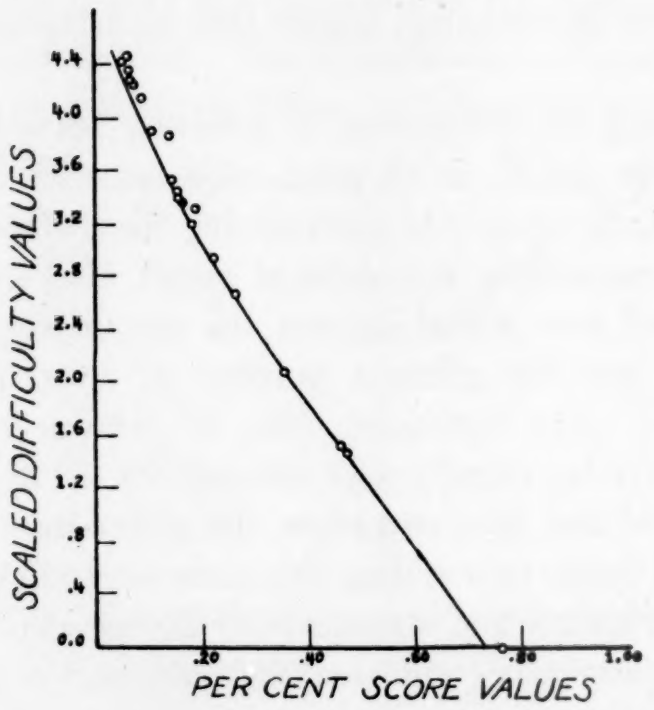

FIG. 15. Scaled means plotted against per cent score means.

The relation between scaled difficulty and length of series is shown to be a negatively accelerated curve. The mean difficulty of recalling a syllable correctly increases sharply as lists are varied in length from three to eighteen syllables by steps of three, but beyond this length, the increase in difficulty, if any, is small and irregular. The relation between difficulty and length of list is rather similar to that existing between mean raw score (taken as the inverse of the percentage correct) and length of list; but as may be seen from inspecting Figures 16 and 17, the relations shown by the two sorts of scores are far from identical. It is quite noticeable that the bend in the curves representing decrease in mean per cent correct with increase in number of syllables is sharper than in those representing increase in difficulty with increase in number of syllables.

TABLE III

	Number of Syllables	Number of Subjects	Mean Per Cent Correct	Absolute Mean Difficulty	Absolute $\sigma_{diff.}$
Immediate Recall	3	62	77.87	0.0000	1.0000
	6	108	46.77	1.5421	.6555
	9	153	35.43	2.0863	.7411
	12	128	26.50	2.6845	.7170
	15	141	22.58	2.9515	.7114
	18	143	18.77	3.2104	.6363
	21	87	16.84	3.3677	.5949
	24	109	15.95	3.4294	.6105
	30	112	16.11	3.4311	.5461
	36	168	14.72	3.5445	.5965
Delayed Recall	3	90	47.55	1.4888	1.2087
	6	167	19.39	3.3167	1.4450
	9	201	14.42	3.8696	1.2656
	12	189	11.03	3.9187	.7633
	15	194	9.08	4.1637	.7334
	18	191	6.49	4.4767	.6625
	21	135	6.62	4.2894	.4659
	24	182	6.48	4.3603	.5763
	30	157	7.49	4.2493	.5534
	36	204	5.52	4.4303	.5160

A problem regarded as more important than the relation between length of series and difficulty is that concerning the effect of a fixed change in conditions at different levels of difficulty. As previously stated, the fixed change used was that constituted by the increase in delay of recall from immediate (or, more exactly, from 116 sec. after the middle of the series) to two days.

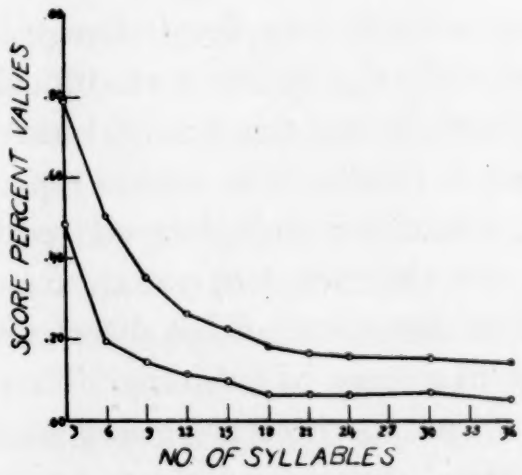

FIG. 16. Mean per cent correct plotted against number of syllables.

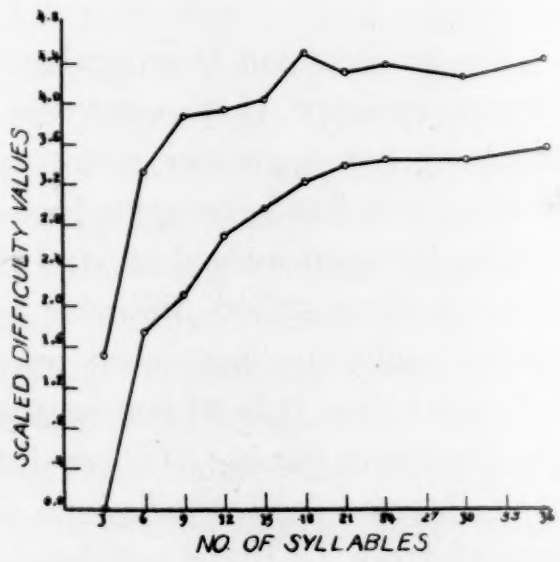

FIG. 17. Scaled means plotted against number of syllables.

The effect of this fixed change at different levels of difficulty produced by varying the length of the lists may be readily noted in Fig. 17 by observing the vertical distance separating the curve showing difficulty for immediate recall from the curve showing difficulty for delayed recall.

On superficial inspection, the curves in Fig. 17 representing increase in mean difficulty with increase in length of list seem roughly parallel. However, when the differences between the means for immediate and delayed recall of lists of the same length are calculated, it becomes evident that as the length of the lists increases, the differences in the means decrease. The differences between the means for immediate and delayed recall are given for each length of list in Table IV.

TABLE IV

THE DIFFERENCE IN ABSOLUTE DIFFICULTY BETWEEN IMMEDIATE AND DELAYED RECALL FOR VARIOUS LENGTHS OF LIST

Number of Syllables	Absolute Mean: Immediate Recall	Absolute Mean: Delayed Recall	Increase in Difficulty Due to Delay
3	0.000	1.489	1.489
6	1.542	3.317	1.775
9	2.086	3.870	1.784
12	2.685	3.919	1.234
15	2.952	4.164	1.212
18	3.210	4.477	1.267
21	3.368	4.289	.921
24	3.429	4.360	.931
30	3.431	4.249	.818
36	3.545	4.430	.885

The progression is irregular, but the trend is definite. In the present instance, then, what was supposed to be a fixed change in difficulty has a greater effect upon difficulty at the less difficult levels. The fixed change referred to is the two-day interval intervening between immediate and delayed recall. The curves representing mean scaled difficulty for immediate and delayed recall most closely approach each other for the two longest, or most difficult, lists. It is of course possible that some of this difference may be due to the use of three different groups of subjects. Thus, in the case of the second group of subjects, who were tested with lists of from 12 to 18 syllables, marked uniformity exists in the effect of delay or difficulty. It is doubtful whether in the case

of any single group of subjects the various lists used with that group were differently affected as regards difficulty by the two day's delay in recall. It seems incredible, however, that differences between the groups of subjects could account for the very marked decrease in effect of delay, from 1.775 units at 6 syllables, to 1.212 units at 15 syllables, and to .818 units at 30 syllables. While admitting that for perfect certainty the matter should be investigated with a single group, the conclusion indicated seems rather clearly to be that under the conditions of this experiment the difficulty of the longer series is much less affected by delay than the difficulty of the shorter series.

It should be pointed out, however, that the change from immediate to delayed recall may not have been a fixed change in difficulty. As previously noted, by immediate recall is here meant recall 116 sec. after the presentation of the middle syllable of the list. Obviously, then, immediate recall meant recall after a pause following the last syllables of over a minute in the case of the shortest lists but after no pause at all in the case of the longest list. On account of the influence of short pauses before recall upon reminiscence (2) it is quite probable that the supposedly fixed change was in reality a variable one. The pause after the last syllable before so-called immediate recall may have enhanced the score in the case of the shorter lists. No such favoring influence was present in the case of immediate recall of the longest lists. Thus, the scores for what has here been termed immediate recall were probably favored in the case of the shorter lists by a factor not present in the case of the longer lists. It follows that the apparent change in the effect of a delay of two days may in reality have been due to this factor of the change in amount of delay before "immediate" recall. It is not always possible to vary one condition without changing another. Thus, in the present instance, it is not clear that keeping constant the pause between the last syllable and recall, at all lengths of series, would obviate the difficulty. For then, of course, the time elapsing between the middle syllable and recall (or in even greater degree, between the first syllable and recall) would vary with length of series. On the whole, it seems improbable that the data here obtained answer

the question of the effect of a fixed change upon difficulty at various levels of difficulty.

Another objective of this experiment was to determine the relation between variability and difficulty under the conditions of the present experiment. Variability is known to vary greatly with degree of difficulty of the task, but the extent and nature of its dependence upon difficulty has not been definitely established. It will be observed from Table III that variability, as expressed by the standard deviation, shows a marked though irregular decrease with increase in length of list, *i.e.,* with increase in difficulty, in both immediate and delayed recall. In other words, the greater the difficulty, the less the variability. However, the results clearly indicate that variability is not solely dependent upon difficulty. It is quite obvious that with difficulty kept constant, variability is much greater with recall after two days than with immediate recall. For example, the absolute mean difficulty of list 6*i* is 1.5421 and that of list 3*d* is 1.4888. The difference in mean difficulty is only .0533, yet the absolute variability of 3*d* is nearly twice that of 6*i* namely, 1.2087 as compared with .6555. Again, there is a difference of only .051 in absolute difficulty between lists 21*i* and 6*d,* while the absolute variability of 6*d* is over twice that of 21*i* (the respective values being 1.4450 for 6*d* and .5949 for 21*i*). Apparently, a two-day delay between presentation and recall increases variability independently of its effect upon difficulty. This result is in accord with Woodrow's (**4**) comment that, though variability varies greatly with the conditions of difficulty of the task, it is probably not solely a function of difficulty, but, in so far as variability is dependent upon difficulty, it seems probable that it decreases with increase in difficulty. The inverse relation here noted between difficulty and variability, both with immediate and delayed recall, seems unquestionably related to Thurstone's finding (**3**) that variability increases with age. The relationship lies in the fact that Thurstone's conclusion means that goodness of performance on a given set of tests shows the greater variation in the case of the group to which the tests offer the less difficulty, namely, the older group. It would be interesting to check this conclusion by giving the memory tests

to younger subjects. From the above statements, it should be expected that the younger subjects would make lower scores, which, if scaled against those of older subjects, would show a smaller age group variability.

Summary.—The task of writing from memory nonsense syllables, each exposed for two seconds, was tested under twenty different sets of conditions of difficulty. These twenty sets of conditions were obtained by using ten different lengths of syllables, extending from three to thirty-six, and both immediate recall and recall after a delay of two days.

The conclusions reached were the following.

(1) The Thurstone-Woodrow absolute scaling technique is applicable to memory data obtained under the conditions of the present experiment. The scaling technique yields a measure of the mean for the group of the difficulty under any given set of conditions of the correct recall of an average syllable.

(2) The relation between mean per cent correct and mean absolute difficulty is an ogive function of a normal distribution curve. If this relation were found to be a general one, raw per cent scores for recall of nonsense syllables could be scaled directly by the use of the ogive formula.

(3) The relation between absolute difficulty and length of list is a negatively accelerated curve in the case of both immediate and delayed recall. Difficulty increases sharply as lists are varied in length from 3 to 18 syllables. With lists of 18 to 36 syllables, the curves flatten out. The curve relating raw (per cent) scores to length of list is also negatively accelerated, but in this curve the relations among the difficulties of different lengths of lists are distorted by a dependence upon unequal units of measurement.

(4) The effect of the change from what is here termed immediate recall to recall after two days is greatest in the case of the shorter lists, that is, the effect of the delay increases as the level of difficulty, on account of increased length of list, decreases. Any interpretation of this conclusion is obscured, however, by two facts. First, since the time-interval between the presentation of the middle syllable of the list and the signal to write was kept constant (at 116 sec.), there occurred a pause, which varied in

length with length of list, between the last syllable of the list and the recall signal. Since this pause was longest in the case of the shortest lists and non-existent in the case of the longest list, the shorter lists may have been favored by the occurrence of reminiscence in the case of immediate recall and thus the apparent effect of two days of delay may have been greatly enhanced as compared with the effect of the same delay in the case of the longest lists. Second, the experiment made use of different groups of subjects for different lengths of lists. For the several lengths of list used with any one of the three groups, there is no evidence of a decrease in the effectiveness of the two days of delay with increase in length of list, but rather a suggestion of constancy in effect.

(5) Absolute variability decreases as difficulty increases, *i.e.*, with increase in length of list, in the case of both immediate and delayed recall. When difficulty is held constant, variability is much greater in the case of delayed recall.

References

1. GLAZE, J. S. The association value of nonsense syllables. *J. genet. Psychol.*, 1928, **5**, 255–269.
2. HOVLAND, C. I. "Reminiscence" following learning by massed and distributed practice. *Psychol. Bull.*, 1936, **33**, 614–615.
3. THURSTONE, L. L. A method of scaling psychological and educational tests. *J. educ. Psychol.*, 1925, **16**, 433–451.
4. WOODROW, H. The measurement of difficulty. *Psychol. Rev.*, 1936, **43**, 341–365.

A FACTOR ANALYSIS OF FORTY CHARACTER TESTS

By

HUBERT E. BROGDEN

Introduction and History.—The aim of this study is to determine just what character traits are involved in the scores from a group of forty tests purporting to measure various phases of character, intelligence, and personality. By a trait is here meant an unidimensional characteristic which varies quantitatively from one individual to another. The tests used are representative of those basic to much of present day theory with respect to character. It is highly desirable, therefore, to determine the traits which are fundamental to the measurements made by these tests. It is believed that the best possible answer to such a question is afforded by factor analysis, since such an analysis should result in a list of the smallest number of independent factors which will explain the obtained test-score intercorrelations, and therefore enable a description of all the test results in the most economical terms obtainable.

Preliminary to discussion of the writer's investigation, a review of previous factor analyses in the field of character will be given. Examination of these previous investigations indicates that much of the work involved use of rating techniques. The well known halo effect is apt to influence markedly the intercorrelations of ratings and, consequently, the factors extracted from the intercorrelations. Devices sometimes used to diminish this effect are of questionable validity.

Apart from this well known defect of ratings, it seems quite probable that many of the interrelations found between ratings are as much a function of the rater's conception of these interrelations as they are a function of the " true " correlations between the characteristics of the persons being rated. Even though the population being rated remained the same, differing factors might

be extracted from ratings by individuals with different racial and cultural backgrounds.

The earliest and among the most thorough of rating studies was that of Webb (23). Webb discovered in the intercorrelations of ratings on forty-eight qualities of character, personality, and intelligence, a factor *w* which was prominently involved in such qualities as tendency not to abandon tasks from mere changeability, tendency not to abandon tasks in the face of obstacles, trustworthiness, and conscientiousness. Spearman's techniques were used in the isolation of this factor. In further analysis of Webb's data, Garnett (5) demonstrated the presence of a second factor *c* which enters into such qualities as cheerfulness, sense of humor, quickness of apprehension, and originality of ideas.

Several more recent studies have followed the trend of Webb and Garnett. Cattell (2), in a rating study which included traits designed to isolate *w* and *c*, discovered a third factor *a*, entering into such qualities as frankness, balance, optimism, generosity and cooperativeness.

Using similar techniques, Studman (20) found, besides *g* and *w*, a new factor, designated *f*, which enters into such qualities as energy, confidence, independence, expressiveness in speech, elated feeling tone, and instability of emotions. Raters were found to be more consistent with each other in rating characteristics dependent upon the *f* factor than in those involving *w*. In his study of twenty-five qualities of school children, Pan-Lin Chi (4) found but one factor. This he assumed to be *w*.

McCloy (12), in analyzing Webb's data by Thurstone's centroid method, isolated much the same factors as from similar data of his own gathered through ratings on thirty-one students. The four factors, found in both sets of data, are a social-anti-social factor, a dominance or positive action-tendency factor, a factor centering in individual qualities as opposed to those which cause an individual to merge with a group, and a factor of positive attitudes. There may be some doubt as to the validity of these factors, since in both studies his rotation procedure gave little that might be described as simple structure.

In data gathered by a detailed and systematic rating method,

Sister Mary McDonough (11) discovered three factors in the forty characteristics studied. The first was evinced by such qualities as looking for sympathy, quarrelsomeness, irritability, and forwardness; the second, will, reliability, generosity, and stability; and the third, cheerfulness, contentment, sympathy, and refinement. A modification of Spearman's techniques was used throughout the study. Maller discovered that the intercorrelations of four composite measures of honesty, coöperation and helpfulness, inhibition, and persistence satisfied the tetrad criterion, and could be accounted for by a single factor. Rating data, carefully checked in many instances by correlation with behavioral observations, was gathered by Berne (1) on fifty pre-school children and later factored by Williams (24) with the use of Thurstone's methods. Williams isolated two factors which were named approach-withdrawal and ascendance-submission. Ryans (17) factored the scores of forty college students on nineteen objective tests. Probably because of too few subjects or the extraction of too few factors, his results were not clear cut. His tests, original with him, were interesting and seem worthy of further investigation.

Procedure.—Eleven sixth grade class-rooms in the Champaign Public School System comprised the tested population. Although girls were included in this population, their tests were not scored since many of the measures, such as knowledge of slang, have different significance for the two sexes. Further omissions resulted in considerable reduction of the population. The results from both a negro and a small rural classroom were omitted. Absences were frequent and many tests were discarded as incomplete or unusable. While the final data were reduced to the scores of one hundred boys, a few of the correlations were computed with still fewer cases. Because of large numbers of name omissions, only ninety-one usable cases remained for tests six and twelve. Ninety-two were available in computing correlations for test number twenty-five, and only eighty-six for test number twenty-three. For all remaining tests, one hundred cases were available.

The effect of the various selective influences just mentioned

was to reduce considerably the heterogeneity of the subjects. Of the ninety-four original subjects from the better sections of the city, seventy-four were retained, while of the fifty-eight from the poorer sections, only twenty-six were finally included. The sample was, then, not representative but selected preponderantly from the higher social strata. Moreover, the subjects discarded, as examination of their scores revealed, tended toward low intelligence and poor character as measured by the tests given. It should be remembered, of course, that this reduction in heterogeneity of the subjects with respect to character and intelligence must have lowered the intercorrelations of the used scores, and, consequently, also lowered the loadings obtained by the factor analysis.

The tests were given in the usual classroom situation, with two exceptions. In order that the two sixth grade classes of one school could be combined, the tests were administered in the basement auditorium. In the second instance, a sliding door separating the two sixth grades was opened, leaving one large room in which the testing was done. The teachers were always present during the testing period but did not interfere except in cases of discipline.

In general, where standardized tests were used, the directions and scoring advocated by the authors were not changed. Usually the directions were printed at the top of each test, and were repeated verbally by the experimenter. There was then allowed a short period of questioning.

Forty measures were used in the final analysis. Thirty of these were performance tests, and ten of them were questionnaires. Of the thirty performance tests, eleven would conventionally be termed measures of intelligence; four, measures of honesty or cheating; three, measures of perseveration; three, measures of persistence; two, measures of variability; two, measures related to slang usage; one, a measure of inhibition; one, a measure of suggestibility; one, a measure of conscientiousness; one, a measure of deportment; and one, a measure of grades in school subjects. The ten questionnaires supposedly test various aspects of character or personality.

Since any factor finally arrived at must be defined, in the first instance, by enumerating the tests which depend heavily upon it, a thorough knowledge of the tests is necessary in order to gain insight into the nature of the various factors. Each test will therefore be described and, where it is thought necessary, the directions and method of scoring stated.

1. Maller Self-Marking Test (9) :—This is a simple achievement test covering such topics as geography, history, and arithmetic. However, mixed with questions of a sixth grade calibre are many far above this level. A key was enclosed and the grading done by the student himself, in order that he might have ample opportunity to cheat. His score is the number of correct answers given by the subject to those questions known to be beyond his ability.

2. Coördination Test (6) :—This well known test measures the tendency of the child to " peep " in order to obtain a good score on a task which should be done with the eyes closed.

3. False Book List (16) :—This test consists of a list of book titles which are to be checked by the child if he wishes to indicate that he has read them. Since many of the titles are fictitious, an overstatement score can be obtained by determining the number of fictitious titles the child claims to have read.

4. Overstatement Test (25) :—In this test the child is asked to state the degree of his knowledge of a number of topics and is then given a test designed to determine the validity of his claims. The score is the difference between the claims of the child and his achievement score.

5. Suggestibility Score:—Originally four suggestibility tests were administered, but since the intercorrelations between these were low, it was decided to combine them into one measure. Two of the tests were forms A and B of the Otis suggestibility test (14). Two new tests devised by Dr. Floyd Ruch were also administered. In the first of these tests, the child is shown many pairs of words, some of which pairs, actually determined by chance, are stated to be extensively used as synonyms. The child is asked to indicate those pairs made up of the words which he would most often use as synonyms. Presumably the suggestible

child would choose the pair stated to be used most extensively. In the second of these tests it is suggested that ink blots resemble objects or animals which they may or may not resemble. Here again the child may accept the suggestion by indicating that he sees the supposed similarity.

6. Persistence in Adding:—An adaption of the Character Education Inquiry's persistence for self and persistence for class test (7) was made in that only one persistence score was used. Thirty groups of two-number additions were administered in consecutive periods of forty-five seconds each. Since adding speed decreases in the later periods, a measure of persistence may be determined from this decrease.

7. Persistence Stories:—An adaption of those used by the Character Education Inquiry (7), this test measures the child's persistence in completing a story which is difficult to read after the point of suspense has been reached. The material is run together, the capitals and small letters are misplaced, and the words are improperly spaced in order to increase the difficulty of reading. As given by Hartshorne and May, a preliminary practice period preceded the actual test. The omission of this practice period in the author's investigation may possibly account for the fact that this test showed low correlations with the other tests.

8. Picture Inhibition Test:—This test, developed by Hartshorne and May (7), was administered with no modifications in the directions or scoring. The score is the difference between the average number of additions completed each period under normal conditions and the average of those finished each period with distractions such as jokes, puzzles, and stories printed directly above the additions.

9. Picture Inhibition Persistence:—A persistence score was derived from the Picture Inhibition Test, by taking the difference between the average of the first two and the average of the last two groups of additions.

10. Slang "A" Score:—The score was that yielded by Schwesinger's (18) tests of knowledge of slang words and slang expressions.

11. Slang "C" Score:—In answering the questions in the

multiple choice division of the slang knowledge test, the subjects could usually make two correct choices, one of which was a slang expression. The "C" score is simply the number of times the subject used a slang expression in making a correct choice.

12. Variability Score:—This measure is the standard deviation of the thirty addition scores obtained for each individual from test number six.

13. Questionable Reading Preferences:—This test, devised by Raubenheimer (**16**), consists of a list of ten book titles, which the child ranks in the order of his preference. It is designed to select children who show interest in reading matter which is judged to be unwholesome.

14. Questionable Character Preferences:—Tests fourteen through nineteen are all taken from Form B-1 of the Opinion Ballots developed by Hartshorne and May (**8**). No changes were made in the directions or scoring.

Test fourteen contains a list of character descriptions of widely differing "goodness". After the child ranks them in the order of his preference, his score is obtained by adding the differences between his ranking and that of competent judges.

15. Opinion Ballot Number Two:—By encircling either *all, most, many, few,* or *no* before a list of statements, the child expresses his opinion about various aspects of authority.

16. Opinion Ballot Numbers Four and Six:—These two ballots were combined to yield one score. In the fourth ballot, the child states whether a list of items such as pull, trickery, a clear conscience, or going to church make for success or failure. In the sixth ballot, he states, in the case of each of a great variety of situations, whether he would be willing to help other individuals.

17. Eighth Opinion Ballot:—This is a questionnaire on various aspects of honest behavior. The child is asked if he would cheat in a variety of situations and is scored by the number of affirmative responses.

18. Ninth Opinion Ballot:—In this test, a series of choices are presented to the subjects. The choices usually involve an opposition between duty and pleasure, or between negligence of duty and pain or discomfort.

19. Stories Test:—This test by Chambers (3) consists of a series of stories attempting to present life-like situations involving ethical choices.

20. Controlled Association:—This test was designed by Raubenheimer (16) to determine the habitual reactions of subjects to social and educational institutions. The subject has to make his choice of the most suitable of several statements concerning a number of such institutions.

21. *S* Perseveration Test:—To obtain the score on this test the speed of writing *S*'s forward, writing them backward, and then writing them alternately forward and backward must be determined. The number completed during the alternation period is then divided by the average of the scores for the other two periods.

22. *V* Perseveration Test:—This test is the same as test number twenty-one except for the use of a different letter.

23. Add-Subtract Perseveration Test:—This test consists merely of a series of two-number additions and subtractions. The letters *A* and *S* over each problem indicate whether the subject should add or subtract.

24. Test of Conscientiousness:—While the Pressey X-O Test (15) was administered, it developed during the scoring that a large percentage had failed to complete all the procedures called for by the directions. Consequently, it was included in the battery not as a test of emotional maturity but as a measure of faithfulness in the carrying out of directions.

25. Deportment Grade:—These were the average of the monthly deportment grades for the first semester.

26. Grades:—The score was the average grade for the first semester.

27. Personality I:—A measure of social adjustment derived from Woodworth's inventory. Weights used in determining this score were the result of a previous unpublished factor analysis made by the writer.

28. Personality II:—This, a second factor found in the Woodworth inventory, seems to be a measure of self-sufficiency.

29. Variability II:—This score was obtained from the Otis

Intelligence Test by calculating the standard deviation of the ten standard scores of the sub-tests.

30–40:—Tests thirty through thirty-nine are the sub-tests of the Otis Intelligence Test. Variable forty is the total Otis score.

Pearson product-moment correlations were computed between the forty scores obtained from the tests just described. These intercorrelations were then subjected to the Thurstone centroid method of multiple factor analysis. After the extraction of eight centroid factors, the residuals were found to be symmetrically distributed around a mean of .009 with a standard deviation of the distribution of .058, which is considerably below the standard deviation of a correlation coefficient of .000 based on 100 cases, that is, .100.

The eight centroid factors were represented as geometric axes and then rotated so as to maximize the number of insignificant loadings and thus minimize the number of factors it is necessary to attribute to any one test. In customary rotation procedure it is attempted to eliminate significant negative loadings. Some of the negative loadings appearing in the final solution here obtained are larger than –.200 and hence may be considered significant. Significant loadings are, of course, to be expected when significant negative correlations are found in the original correlational matrix. However, of the five tests showing negative loading larger than —.200, that is, Tests 2, 11, 21, 22 and 31, two of them, 22 and 31, do not show significant negative correlation with the other tests. These latter two tests, however, show negative correlations between —.100 and —.200, and this fact, together with the influence of chance factors upon the correlations, may have necessitated negative loadings. It should be mentioned, though, that a number of other variables had negative correlations between —.100 and —.200, but showed no significant negative loadings after rotation of axes; and these negative correlations may therefore be regarded as due to chance errors.[1]

Results.—The rotated factorial matrix is presented as Table I. Each column shows the loadings of the forty tests with one of the eight factors. The column headed h^2 shows the sum of the

[1] On this point, see (13).

squares of the factor loadings of each test and therefore that proportion of the total variance of the scores of a test which is

TABLE I

THE ROTATED FACTORIAL MATRIX

	I	II	III	IV	V	VI	VII	VIII	h^2
1.	.178	.586	.202	.136	.092	.040	.041	—.104	.419
2.	—.212	.626	.042	.149	.077	.217	.088	—.222	.517
3.	—.196	.547	.047	—.034	.145	.051	—.033	.274	.441
4.	—.080	.529	—.035	.295	.198	.089	—.031	.068	.427
5.	.534	—.125	.150	.202	.113	—.054	—.044	.075	.387
6.	.289	—.036	.532	.077	.064	.207	—.079	—.168	.455
7.	—.199	.042	.169	.328	—.031	.170	—.016	.225	.259
8.	.011	—.041	.495	—.027	—.012	—.061	.210	.111	.308
9.	—.016	—.056	.581	—.091	.086	.059	.063	.154	.388
10.	.004	.010	.105	.404	.603	—.103	.012	—.139	.568
11.	.147	.028	—.057	—.036	—.282	.547	.056	.077	.415
12.	.218	.038	.524	.016	.033	.055	.055	—.106	.342
13.	—.055	.028	—.039	—.130	.356	.440	—.034	—.060	.347
14.	.306	.342	.254	.304	.175	.047	.057	.192	.470
15.	.216	.008	.189	.306	.343	.228	.302	.105	.444
16.	.036	.120	.211	.115	.354	.004	.411	—.045	.370
17.	.010	.309	—.014	.009	.110	—.083	.556	—.035	.425
18.	.104	.227	.050	.161	.473	.506	.025	.125	.587
19.	.013	—.026	.194	.122	.438	.186	.322	.342	.500
20.	.088	.147	.085	.351	.151	.125	.237	.478	.483
21.	—.456	.178	.271	.141	.020	—.160	.049	.242	.420
22.	—.377	.256	.333	.021	.122	—.117	—.167	.053	.350
23.	.046	.136	.357	.084	.080	—.111	—.040	.424	.355
24.	.373	.185	.256	.269	—.050	—.076	—.050	.102	.332
25.	.000	.145	.049	—.002	.052	.116	.263	.359	.237
26.	—.044	.169	.371	.371	.404	—.013	.077	.087	.484
27.	.057	.074	—.022	.443	.216	.041	.438	.148	.468
28.	—.194	.002	.198	.259	—.179	.405	—.065	.125	.360
29.	.155	—.168	.122	.330	.062	.133	.067	.118	.216
30.	.153	—.112	—.089	.404	.554	.019	—.055	.142	.538
31.	.066	.046	—.038	.416	.345	—.117	—.223	.227	.415
32.	.119	.062	.139	.597	.425	.060	.271	—.077	.657
33.	—.072	.216	.026	.628	.356	—.099	—.031	.105	.594
34.	.130	.094	—.041	.708	.094	.038	.141	.011	.559
35.	.049	—.030	.008	.739	.017	—.073	—.035	—.047	.559
36.	.053	.157	.161	.524	.181	.204	—.032	.496	.653
37.	—.103	.108	.079	.743	.035	.020	.062	—.053	.580
38.	—.057	.037	—.132	.642	.137	.318	.100	.003	.564
39.	—.063	—.196	.214	.565	.213	.047	.254	.255	.585
40.	.113	—.061	.014	.853	.386	—.024	.094	.304	.995

accounted for by variation in the amounts possessed of the various factors by the individuals composing the group.[2]

Each factor will be described first of all by simply listing the tests which show loadings with it of .300 or higher. An attempt

[2] Mimeographed copies of the table of original correlations, the unrotated centroid factor loadings, and the transformation matrix will gladly be furnished upon request addressed to the author.

will then be made to state the nature of the element common to the tests having high loadings with the given factor. These statements should be considered theoretical until their scientific value has been demonstrated by enabling construction of new tests with high loadings in the factor under discussion. Ideally, the nature of all the factors should be so well specified that one could at will devise further tests having either high or insignificant loadings with any of the factors.

This factor bears considerable resemblance to Webb's *w* factor. This resemblance is conspicuous in the high loadings with this factor shown by the scores of test twenty-four (conscientious-

FACTOR I

5.	Resistance to Suggestion	.534
24.	Conscientiousness	.373
14.	Questionable Character Preferences	.306
21.	*S* Perseveration	—.456
22.	*V* Perseveration	—.337

ness) and tests twenty-one and twenty-two (tests of perseveration). That tests of perseveration heavily involve *w* has been claimed by a number of previous investigators. The loadings of Test 14, Questionable Character Preferences, with this factor may be accounted for on the grounds that the difference between the good and bad characters portrayed in the test is largely a matter of such qualities as trustworthiness, conscientiousness, and tendency not to abandon tasks from mere changeability—all characteristics with high loadings in *w*. There is no evidence from previous studies concerning the degree of relationship between suggestibility and *w*. The high loading of the suggestibility score in Factor I might, then, be presented as new evidence concerning the nature of *w*.

Despite these observed similarities, any identification of Factor I with *w* should be made with extreme caution. This is necessitated by the contrast between Webb's study and the present one in both the statistical methods of analysis and the means of collecting the data. Furthermore, the vagueness, both of the names applied to the traits which were rated in the isolation of *w* and the names applied to the tests in the present study, should make such an identification even more dubious.

Factor II

1.	Maller Honesty Test	.586
2.	Cördination Test	.626
3.	False Books List	.547
4.	Overstatement Test	.529
14.	Questionable Character Preferences	.342
17.	Opinion Ballot Number Five	.309

Since the first four tests measure either tendency to cheat or tendency to misstate accomplishments, this factor may be termed an honesty factor. It is interesting to note that verbal attitude toward cheating, as measured by Opinion Ballot Five, is related to the overt behavior measured by Tests 1 to 4, inclusive. Improvement could probably be made in both tests fourteen and seventeen as measures of this factor. To accomplish this, one might begin by selecting two groups of individuals, namely, those possessing relatively large amounts of Factor II and those possessing relatively small amounts of Factor II; and then, by determining which items in Test 17 best differentiate these two groups, one could ascertain which items of Test 17 were the best measures of Factor II. On the basis of such knowledge, it should be possible to revise Test 17 so as to make it correlate better with the factor. This same technique could be used to improve test fourteen as a measure of this factor. Moreover, by thus determining the items in Tests 14 and 17 which best measure this factor, one might hope to obtain an improved understanding of its nature.

Test eight involves resistance to distractions in the nature of jokes, pictures and puzzles; most of the other tests seem to call for resistance to either boredom or fatigue. The main attribute of the person high upon this factor may, therefore, be a tendency to continue a steady work output in spite of the distractions

Factor III

6.	Persistence in Adding	.532
8.	Picture Inhibition	.495
9.	Picture Inhibition Persistence	.581
12.	Non-variability in Adding	.524
26.	Grades	.371
23.	Add-subtract Perseveration Test	.357
22.	*V* Perseveration Test	.333

arising from fatigue, boredom, jokes, puzzles, or other sources which might cause variations from one minute to the next. The term persistence is tentatively adopted. Although the highest loadings all involve addition, the loadings of Tests 26 and 22 indicate that the influence of this factor is not limited to adding activity.

Factor IV

It will be helpful, before attempting to interpret Factor IV, to discuss its relation to Factor V and to mention a problem which arose during rotation. On the plane picturing these two factors, two poorly defined but correlated (non-orthogonal) clusters of tests were so located inside the positive quadrant that, if a rotation were made to produce loadings of nearly zero in the case of one of these clusters on one factor, the remaining cluster would possess sizable loadings upon both Factors IV and V. This situation makes an interpretation of these factors rather difficult, and, indeed, throws doubt upon the proper position into which to rotate the axes. It seemed that possibly the most satisfactory result would be attained by removing the restriction of orthogonality of axes, since axes could then be passed through both of the aforementioned clusters. This procedure, which left the orthogonal Factor V loadings unchanged, gave new Factor IV loadings, which are listed below in the column headed "Oblique".

		Oblique	Orthogonal
40.	Otis Intelligence Total Score	.579	.853
37.	Directions	.643	.743
35.	Similarities	.648	.739
34.	Analogies	.585	.708
33.	Arithmetic Problems	.393	.628
38.	Geometric Figures	.506	.642
32.	Proverbs	.334	.597
39.	Narrative Completion	.403	.565
36.	Memory	.381	.524
27.	Personality I	.293	.443
31.	Disarranged Sentences	.210	.416
30.	Vocabulary	.103	.404
10.	Slang "A" Score	.080	.404
26.	Grades	.143	.371
20.	Controlled Association	.242	.351
29.	Non-variability II	.264	.330
7.	Persistence Stories	.306	.328
15.	Opinion Ballot Number Two	.103	.306
14.	Questionable Character Preferences	.189	.304

The writer does not feel that the data justify any final conclusions as to the nature of Factors IV and V. Factor IV, it is true, shows its highest loadings (in the case of both the oblique and orthogonal solutions) with tests one would expect to be heavily loaded with Spearman's *g* factor. The centroid method is hardly the proper one to use, however, if one is to determine loadings with *g*. Another possibility is that Factor IV, either oblique or orthogonal, is identical with Thurstone's verbal relations factor *V*. Both Factors IV and Thurstone's *V* show similar loadings in tests having to do with proverbs, verbal analogies and vocabulary—the only tests used in the present study similar to tests used by Thurstone.

Factor V

10.	Slang "A" Score	.603
30.	Vocabulary	.554
18.	Opinion Ballot Nine	.473
19.	Stories Test	.438
32.	Proverbs Test (from Otis)	.425
26.	Grades	.404
40.	The Otis Intelligence Score	.386
13.	Reading Preferences	.356
33.	Arithmetic Problems	.356
16.	Opinion Ballots Three and Four	.354
31.	Disarranged Sentences	.345
15.	Opinion Ballot Two	.343

It is suggested that grades, the Slang "A" score, the vocabulary test, the arithmetic problems test, and possibly the proverbs test, all, to some degree, measure achievement, and that Factor V has to do with achievement. Since several attitude questionnaires have high loadings upon this factor, these must also be considered in a discussion of its nature. It seems possible that these questionnaires measure attitudes involved in the learning or achievement of the child. An hypothesis as to the nature of this factor, then, stated as concisely as possible with the evidence at hand, is that Factor V is a certain component *x* which is common to both achievement and certain attitude measures.

For this factor, the terms self-control, inhibition, or, possibly, dutifulness are suggested by examination of the items of Opinion Ballot Nine. This interpretation seems supported by the presence

FACTOR VI

11.	Slang "C" Score	.547
18.	Opinion Ballot Nine	.506
13.	Questionable Reading Preferences	.440
28.	Personality Measure II	.405
38.	Geometric Figures Test	.318

of Tests 11 and 13, which measure respectively the tendency not to choose slang words or phrases in responding to multiple choice tests of slang knowledge and the tendency not to choose as possible reading material titles suggesting dime novels. While the meaning of Personality Measure II is somewhat doubtful, it seems possibly to be a measure of self-sufficiency, and hence not inconsistent with the above interpretation of Factor VI. According to this interpretation the loading of .318 with Geometric Figures would probably have to be regarded as insignificant. Determination of the particular items of the attitude questionnaires (Test 18) which are good measures of this factor should help to clarify the interpretation.

The child obtaining a high score on this factor feels that boys and girls should act honestly in all situations, is socially well adjusted as measured by a questionnaire, feels that boys and girls should be helpful in most situations, believes that the "good" will succeed, claims to act in the approved manner in the life-like situations described in test nineteen, and responds verbally as though he were well adjusted in the home and school. The picture seems to be that of the child who accepts without question the codes and mores of parents, schools, and churches. This interpretation does not imply that the children who accept these codes do not violate them. Possibly the main characteristic of the child who stands high with respect to this factor is the absence of conflict with the various institutions enforcing or teaching the

FACTOR VII

17.	Opinion Ballot Eight	.556
27.	Personality I	.438
16.	Opinion Ballots Four and Six	.411
15.	Opinion Ballot Two	.302
19.	Stories	.322

social code. If such were the nature of this factor, the loading of the social adjustment score, Test 27, suggests an interesting extension to the meaning of this factor as well as an aid to our understanding of social adjustment.

Factor VIII

One or more meaningless factors are to be expected in any thorough analysis by the centroid method. Factor VIII appears to be such a factor.

Summary.—It is concluded, then, that of the eight factors involved in the forty tests of this study, five pertain almost exclusively to the character tests and may therefore be referred to as character factors. None of the ten Otis intelligence tests showed significant loadings with any of these five factors. One factor (IV), on the other hand, appears to be primarily an intellectual factor, since the nine scores showing the highest loadings with it are all Otis intelligence test scores. This factor resembles both Spearman's g factor and Thurstone's verbal relations factor V. Another factor (V) was found which showed loadings with both intelligence and attitude tests. Concerning the identification of this factor, it can only be suggested that it has something to do with both achievement and attitudes.

The five character factors include an honesty factor (II), a persistence factor (III), a factor tentatively identified as the w factor of the Spearman school (I), a self-control factor (VI), and an "acceptance of the moral code" factor (VII).

Bibliography

1. Berne, Esther Van Cleave. An experimental investigation of social behavior patterns in young children. *Univ. Ia. Stud. Child Welf.*, 1930, Vol. IV, No. 3. Pp. 93.
2. Cattell, R. B. Temperament tests. I. Temperament. *Brit. J. Psychol.*, 1932, 23, 308–329.
3. Chambers, Edward V. A study of dishonesty among the students of a parochial secondary school. *Ped. Sem.*, 1926, 33, 717–728.
4. Chi, Pan-Lin. Statistical analysis of personality rating. *J. exper. Educ.*, 1937, 5, 229–245.
5. Garnett, J. C. M. General ability, cleverness, and purpose. *Brit. J. Psychol.*, 1917–1919, 9, 345–366.
6. Hartshorne, H., and May, M. A. Studies in the nature of character. I: Studies in deceit. New York: Macmillan, 1930. Pp. 414.

7. HARTSHORNE, H., MAY, M. A., and MALLER, J. B. Studies in the nature of character. II: Studies in service and self control. New York: Macmillan, 1930. Pp. 559.
8. HARTSHORNE, H., MAY, M. A., and SHUTTLEWORTH, F. K. Studies in the nature of character. III: Studies in the organization of character. New York: Macmillan, 1930. Pp. 503.
9. MALLER, J. B. The case inventory: for measurement of some fundamental aspects of character and personality. New York: Teach. Coll., Columbia Univ., 1935.
10. MALLER, J. B. General and specific factors in character. *J. soc. Psychol.*, 1934, **5**, 97–102.
11. MCDONOUGH, SISTER MARY. The empirical study of character. Washington: Catholic Univ., 1929. Pp. 222.
12. MCCLOY, C. H. A factor analysis of personality traits to underlie character education. *J. educ. Psychol.*, 1936, **27**, 375–387.
13. MOSIER, C. I. Influence of chance error on simple structure: an empirical investigation of the effect of chance error and estimated communalities on simple structure in factoral analysis. *Psychometrika*, 1939, **4**, 33–44.
14. OTIS, MARGARET. A study of suggestibility of children. *Arch. Psychol.*, 1924–1925, Series 2, **11**, No. 70. Pp. 108.
15. PRESSEY, S. L. A group scale for investigating the emotions. *J. abnorm. (soc.) Psychol.*, 1921, **16**, 55–64.
16. RAUBENHEIMER, A. S. An experimental study of some behavior traits of the potentially delinquent boy. *Psychol. Monogr.*, 1925, **34**, No. 6. Pp. 106.
17. RYANS, DAVID G. An experimental attempt to analyze persistent behavior. *J. gen. Psychol.*, 1938, **19**, 333–353.
18. SCHWESINGER, C. C. Slang as indication of character. *J. appl. Psychol.*, 1926, **10**, 245–263.
19. SPEARMAN, C. E. The abilities of man. New York: Macmillan, 1927. Pp. 415.
20. STUDMAN, L. G. Studies in experimental psychiatry. V: "W" and "F" factors in relation to traits in personality. *J. ment. Sci.*, 1935, **81**, 107–137.
21. TERMAN, L. M. Mental and physical traits of a thousand gifted children. I: Genetic studies of genius. Stanford University: Univ. Press, 1925. Pp. 648.
22. THURSTONE, L. L. The vectors of mind. Chicago: Univ. Chicago Press, 1935. Pp. 266.
23. WEBB, EDWARD. Character and intelligence. *Brit. J. Psychol. Monogr. Suppl.*, No. 3. Pp. 99.
24. WILLIAMS, H. M. A factor analysis of Berne's "Social behavior patterns in young children." *J. exper. Educ.*, 1935–1936, **4**, 142–146.

THE EFFECT OF PRACTICE UPON STANDARD ERRORS OF ESTIMATE

By

LELAND P. BRADFORD[1]

Problem.—The original purpose of this study was to determine the effect of practice upon trait variability in the case of seven memory tests. By trait variability is here meant the variability or unevenness in the amounts of several traits possessed by a single individual. It is a term which should be sharply distinguished from individual variability, by which is meant the variability in the amounts of one specified trait possessed by the individuals composing a group. Difficulties mentioned in the latter part of this paper seem to make it extremely difficult, if not impossible, satisfactorily to answer this problem. The data obtained, therefore, will be presented for their bearing upon problems related to, but somewhat different from, the problem originally contemplated. The chief of these problems are the following:

(1) What is the effect of practice on test intercorrelations? The answer to this question must be ascertained before the answer to the following one can be determined.

(2) Does practice in each of several tests make it possible to estimate scores in one test from those made in another test more accurately after practice than before practice? Statistically stated, are the standard errors of estimate reduced by practice?

Tests.—The tests used were the following: (1) Digit and Letter Squares Recall; (2) Nonsense Syllable Recall; (3) Nonsense Syllable Recognition; (4) Digit Recall; (5) Digit Recognition; (6) Sanskrit-English Vocabulary; and (7) Prose Selection. A brief description of each test follows.

1. Digit and Letter Squares Recall. Five rows of four squares each, containing a letter or a number in each square, were presented on a stereopticon slide for an exposure time of 20 sec. Immediately after the presentation of the

[1] This investigation was initiated under the direction of the late Professor Edward H. Cameron.

slide, the subjects attempted to write as many as possible of the numbers and letters in the correct squares of a mimeographed blank. The score was the number of correct numbers or letters placed in the correct squares.

2. Nonsense Syllable Recall. Sixteen three letter nonsense syllables were presented in a single column on a lantern slide for 24 sec. After presentation, the subjects wrote the recalled syllables. One point was scored for each *letter* reproduced in the proper position.

3. Nonsense Syllable Recognition. The sixteen syllables presented in the previous test were mixed with forty-eight other nonsense syllables and presented to the subjects on mimeographed sheets. The syllables were listed in three columns. The letters *Y* (yes) and *N* (no) appeared after each syllable. The subjects were to circle *Y* if they had just previously seen the syllable, and *N* if they had not. The time allowed was 3 min. Since the subjects were told that, of the sixty-four syllables on the sheet, sixteen had been seen previously, there were three chances of guessing wrong to one of guessing right. Consequently, the scoring formula used was Right—(1/3 Wrong).

4. Digit Recall. Twelve four-place numbers were presented in one column by a lantern slide for 48 sec. Subjects were tested for immediate recall. The score was the number of digits in the right place.

5. Digit Recognition. This test was similar to the Nonsense Syllable Recognition Test. The formula for scoring was Right—(1/4 Wrong).

6. Sanskrit-English Vocabulary. Thirty Sanskrit words with English meanings were presented for study on a mimeographed sheet for 2½ min. The subjects were then given a second sheet containing the Sanskrit words arranged in a different order. They were instructed to recall as many English meanings as possible. The time allowed was 5 min. One point was scored for each English meaning correctly recalled.

7. Prose Selection. This test consisted of a selection of non-fictional prose material containing one hundred ideas and approximately two hundred words. An idea was defined, in this instance, as being a noun or a qualifying adjective. Verbs and verb phrases were not considered as separate ideas. The material composing the selections was taken from a number of books dealing with science, drama, and music. The selections were judged to be of a relatively high level of difficulty. They were presented on a mimeographed sheet with an exposure time of 90 sec. Following the reading of the selection, the subjects were instructed to turn the sheet over and rewrite the selection in their own words. The time allowed was 8 min. Each idea recalled was scored as one point.

Procedure.—On the first day the subjects were given only the necessary directions in order to take the tests. Before testing on the second day, the subjects were told that the individual making the largest total percentage of improvement over his previous day's scores would receive a ticket to one of the local theaters, and that one ticket would be given each day. The subjects were also informed that two grand prizes would be given at the end of the practice periods, although they were not informed as to what the grand prizes would be. The first prize was awarded for an accumulation of points based upon daily and weekly improve-

ment. The second prize was given to that individual who held the most records, or highest scores, made in each test.

The theater tickets were not always given to the individual making the greatest improvement. This incentive was controlled so that one individual would win no more than one ticket. Otherwise some individuals might become discouraged and the incentive lose its effectiveness. Also, for the first grand prize, the subjects were kept informed of what purported to be their standing in points. Each subject was told once a week that his points were within striking distance of the leader. Every day each subject was shown his score in each test on the previous day.

In order to determine the effect of age differences upon the problems studied, two groups were used. Experiment I was conducted with forty-three university juniors and seniors, and Experiment II, with thirty-four tenth grade students in the University High School. Twenty-five forms of each of seven memory tests were administered to each group. The tests were made on twenty-five successive days (with the exception of Saturdays, Sundays, and holidays, and, in Experiment II, a ten-day interlude occasioned by the Christmas vacation), and a different test form was used each day.

In the case of each test, scores on the second and third days were added together to form what are termed initial scores. The scores made on the first day were not included, as testing and motivational conditions are necessarily different on this day from those on the following days. Scores on the twenty-fourth and twenty-fifth days were added together to form the final scores. In the case of the tenth grade group (Experiment II), intermediate scores were also calculated by adding the scores made on the two days immediately preceding the Christmas vacation (the 16th and 17th days) and also those made on the two days immediately following the vacation (the 18th and 19th days).

Results.—Since the amount of practice given was rather small, the first data to be presented will indicate just how effective was this practice in bringing about an improvement in mean score. Table I shows this increase in mean score for each test in the case of each group, both as a gain in raw score and as a percentage of

initial score. It will be noticed that the per cents of gain for each test are very similar in the two experiments. Only in the Prose Selection Test was the difference in per cents of gain between the two experiments at all large.

The further data to be presented consist of the following calculated values: (1) the test intercorrelations; (2) the standard errors in the estimation of one set of test scores from another. In each case, the effect of practice is made clear by presenting the values calculated from both initial and final scores. In addition (3), it will be explained why the standard deviations of standard scores cannot be satisfactorily used to determine the effect of practice on trait variability.

TABLE I

INCREASE IN MEAN SCORE PRODUCED BY PRACTICE

Test	Initial	Final	Gain	% Gain
Experiment I				
1. Squares	11.4	19.2	7.8	68
2. Syllables, Recall	27.0	38.4	11.4	42
3. Syllables, Recognition	85.2	96.9	11.7	13
4. Digits, Recall	23.6	35.2	11.6	49
5. Digits, Recognition	75.4	85.7	10.9	13
6. Sanskrit-English	15.8	24.7	8.9	56
7. Prose Selection	53.7	101.9	48.2	90
Experiment II				
1. Squares	12.0	20.2	8.2	68
2. Syllables, Recall	23.2	33.9	10.7	46
3. Syllables, Recognition	81.2	94.6	13.5	16
4. Digits, Recall	20.3	27.9	7.6	37
5. Digits, Recognition	78.5	82.4	3.9	05
6. Sanskrit-English	14.8	23.4	8.6	57
7. Prose Selection	42.9	70.3	27.4	64

The effect of practice on test intercorrelations.—The test intercorrelations are needed for the calculation of the standard errors of measurement, but the effect of practice on test intercorrelations is itself a topic of some interest. The product-moment coefficients of correlation are shown in Tables II and III. Each test number in the top lines of the tables refers to the same test as in the case of Table I. The reliability coefficients, on the line labeled "R", are those yielded by the Spearman-Brown correction applied to the coefficient of correlation between the two

successive days, the scores of which were added to form the scores in question. The number of subjects was forty-three in the case of the university students (Experiment I) and thirty-four in the case of the high school students (Experiment II). Below each

TABLE II

TEST INTERCORRELATIONS (EXPERIMENT I)

Initial Intercorrelations

Test	1	2	3	4	5	6	7
R.	.62	.76	.85	.77	.86	.86	.84
Mean	11.4	27.0	85.2	23.6	75.4	15.8	53.7
$\sigma_{dis.}$	3.3	6.2	7.4	5.6	9.6	6.0	11.3
2.	.269						
	.392						
3.	.219	.326					
	.302	*.405*					
4.	.255	.230	.041				
	.369	*.301*	*.050*				
5.	.388	.097	.405	.120			
	.532	*.120*	*.473*	*.147*			
6.	.332	.283	—.049	.219	—.024		
	.454	*.349*	*—.057*	*.269*	*—.028*		
7.	—.034	.071	—.014	—.024	.035	.119	
	—.047	*.089*	*—.016*	*—.030*	*.041*	*.140*	
Mean	.238	.212	.155	.140	.170	.143	.025
	.333	*.276*	*.193*	*.183*	*.214*	*.184*	*.029*

Final Intercorrelations

Test	1	2	3	4	5	6	7
R.	.76	.90	.93	.93	.86	.94	.95
Mean	19.2	38.4	96.9	35.2	85.7	24.7	101.9
$\sigma_{dis.}$	3.9	6.8	8.4	7.5	6.1	6.4	11.6
2.	.285						
	.344						
3.	.031	.243					
	.036	*.265*					
4.	.334	.491	.318				
	.397	*.536*	*.342*				
5.	.404	.335	.557	.329			
	.499	*.380*	*.619*	*.368*			
6.	.030	.508	.252	.296	.212		
	.035	*.552*	*.269*	*.316*	*.236*		
7.	—.039	.346	.411	.295	—.011	.302	
	—.046	*.374*	*.437*	*.314*	*—.012*	*.319*	
Mean	.174	.368	.306	.344	.304	.266	.217
	.211	*.408*	*.328*	*.379*	*.348*	*.288*	*.231*

coefficient is given in italics the same coefficient after correcting for attenuation by dividing the uncorrected coefficient by the square root of the product of the Spearman-Brown reliability coefficients of the two tests concerned.

Table II indicates that in the case of Experiment I, with university students, fifteen correlations increased and six decreased with practice. Five of the six decreases involved the same test, namely, Test 1, the Digit and Letter Squares Test. Table III, pertaining to the effect of practice in the case of the high school students, shows that twenty of the correlations increased whereas only one of them decreased. There appear to be two possible explanations of the much greater preponderance of increases in correlation in the case of the high school group. One explanation of the difference between the two groups might be that it is simply an age difference. Woodrow (3) has found, however, that wide differences between groups all of the university level are shown with respect to the effect of practice in test intercorrelations. In one group, for example, he found nine increases and only one decrease in test intercorrelations, while in another group, he found fifteen decreases and only six increases. A second but more plausible explanation of the difference between the high school and university groups observed in the present study is that the greater increase in test intercorrelations noted in the case of the high school students was due to the fact that a number of the high school students lost interest in the experiment and that lack of perseverance on the part of these students lowered their scores in all tests below what they would otherwise have been. Such a lack of perseverance on the part of a number of the students would augment a tendency towards an increase with practice in the variability of the group; and increase in variability produced in this manner would cause the correlations to increase greatly. That the variability of the high school group increased with practice far more than the university group is obvious from the values of $\sigma_{dis.}$ given in Tables II and III. The increases in these values for the university group, in order, for Tests 1 to 7, were as follows: .6, .6, .9, 2.0, −3.5, .4, and .4, while for the high school group they were .7, 2.9, .8, 3.9, 3.6, 2.0, and 5.9. The average increase in the case of the university group was only .2, while in the case of the high school group it was 2.8, or fourteen times as great as in the case of the university group, in spite of the fact that the two groups did not differ greatly in the size of

the initial standard deviations. In view of the customary increase in variability with practice, it is interesting to note that in the case of Test No. 5, Digit Recognition, the university group showed a large *decrease* with practice in group variability, the $\sigma_{dis.}$ dropping from 9.6 to 6.1, in spite of an increase in mean score from 75.4 to 85.7. In the case of the same test, the variability of the high school group *increased* from 6.4 to 10.0, while the mean score increased less than in the university group—from 78.5 to 82.4.

TABLE III

TEST INTERCORRELATIONS (EXPERIMENT II)

Initial Intercorrelations

Test	1	2	3	4	5	6	7
R.	.78	.83	.80	.85	.78	.95	.91
Mean	12.0	23.2	81.2	20.3	78.5	14.8	42.9
$\sigma_{dis.}$	3.0	5.8	9.7	4.6	6.4	7.0	13.7
2.	.221						
	.274						
3.	.262	.321					
	.331	*.393*					
4.	.509	.624	.052				
	.624	*.742*	*.063*				
5.	—.116	.049	.222	.024			
	—.148	*.061*	*.278*	*.029*			
6.	.489	.498	.349	.532	.181		
	.568	*.559*	*.399*	*.591*	*.209*		
7.	.316	.325	.480	.393	.176	.620	
	.375	*.373*	*.563*	*.446*	*.208*	*.666*	
Mean	.280	.339	.281	.355	.089	.445	.385
	.336	*.400*	*.338*	*.416*	*.106*	*.498*	*.438*

Intermediate Intercorrelations (16th and 17th days)

Test	1	2	3	4	5	6	7
R.	.83	.88	.82	.90	.86	.96	.90
Mean	17.0	29.3	88.7	23.7	77.8	19.5	62.3
$\sigma_{dis.}$	4.3	7.0	9.1	6.5	9.7	8.3	17.4
2.	.441						
	.515						
3.	.158	.433					
	.179	*.509*					
4.	.593	.653	.422				
	.685	*.733*	*.490*				
5.	.117	.271	.690	.354			
	.138	*.311*	*.821*	*.413*			
6.	.630	.610	.434	.590	.419		
	.703	*.663*	*.487*	*.634*	*.460*		
7.	.604	.541	.308	.563	.208	.764	
	.698	*.607*	*.358*	*.625*	*.236*	*.821*	
Mean	.424	.491	.407	.529	.343	.574	.498
	.486	*.556*	*.474*	*.596*	*.396*	*.628*	*.537*

TABLE III—*Continued*

Intermediate Intercorrelations (18th and 19th days)

Test	1	2	3	4	5	6	7
R.	.77	.89	.82	.88	.88	.94	.98
Mean	16.7	29.7	90.7	25.5	78.6	19.9	64.6
$\sigma_{dis.}$	4.2	7.8	11.1	7.5	8.7	8.7	19.2
2.	.639						
	.770						
3.	.266	.679					
	.334	*.793*					
4.	.582	.690	.446				
	.705	*.779*	*.518*				
5.	.253	.382	.354	.487			
	.306	*.431*	*.411*	*.553*			
6.	.615	.814	.631	.674	.388		
	.719	*.889*	*.717*	*.740*	*.426*		
7.	.519	.630	.546	.623	.221	.770	
	.593	*.673*	*.606*	*.670*	*.237*	*.802*	
Mean	.479	.639	.487	.583	.347	.648	.551
	.571	*.722*	*.563*	*.661*	*.394*	*.715*	*.597*

Final Intercorrelations

Test	1	2	3	4	5	6	7
R.	.81	.95	.87	.96	.94	.97	.98
Mean	20.2	33.9	94.6	27.9	82.4	23.4	70.3
$\sigma_{dis.}$	3.7	8.6	10.5	8.5	10.0	9.1	19.6
2.	.470						
	.534						
3.	.482	.631					
	.573	*.693*					
4.	.532	.717	.446				
	.601	*.751*	*.487*				
5.	.257	.640	.527	.559			
	.293	*.677*	*.571*	*.588*			
6.	.717	.739	.713	.678	.555		
	.805	*.770*	*.775*	*.702*	*.581*		
7.	.569	.635	.767	.508	.454	.789	
	.635	*.658*	*.829*	*.523*	*.473*	*.809*	
Mean	.504	.638	.595	.573	.498	.698	.629
	.573	*.680*	*.654*	*.608*	*.530*	*.740*	*.654*

Objective evidence that the much greater increase in variability shown by the high school students was due to lack of persistence on the part of a portion of a group—believed to be the case from general observation—was secured by a study based on ratings by four members of the high school faculty. Two of the raters were men and two were women. One was an administrative officer of the school, while the others were each teachers of different subjects, English, Mathematics, and Science. Each rater thus had the opportunity to observe the students under different conditions from those under which he was observed by any of the

others. The students were rated on a five point scale for what purported to be three different traits: (1) perseverance in tasks after they became monotonous; (2) ability to resist the influence of lack of interest on the part of other students; and (3) likelihood of acceptance of the seriousness of an experiment such as the one performed. The average correlations between pairs of raters on each of the traits ranged from .71 to .90 and averaged .79, so that it is evident that the raters showed a high degree of consistency among themselves. That the three traits rated were very similar in the minds of the raters, as was to be expected, is indicated by the high average intercorrelation of the ratings,

TABLE IV

CORRELATIONS BETWEEN RATINGS OF PERSEVERANCE AND INITIAL AND FINAL MEMORY SCORES

Tests	Initial	Final
1. Squares	.279	.445
2. Syllable Recall	.347	.533
3. Syllable Recognition	.114	.500
4. Digit Recall	.448	.452
5. Digit Recognition	—.009	.404
6. Sanskrit-English	.491	.651
7. Prose Selection	.482	.729

namely, .89. In view of this high intercorrelation of the three rated traits, all ratings of all judges were averaged to obtain a persistence or perseverance score, and these average perseverance scores were correlated with both initial and final memory test scores. The raters of course knew nothing about the test scores. Yet the final memory test scores, in the case of all tests, correlated higher, in some cases far higher, than the initial memory test scores with the perseverance ratings. The correlations are given in Table IV.

Mere increase with practice in the variability of the memory tests would not explain the increase in the correlations with persistence ratings shown by Table IV. This latter result can best be explained on the assumption that lack of persistence on the part of some students did in fact lower the scores of those students, and thus brought about an increase both in the correlations with the ratings and the intercorrelations of the scores of

the memory tests. These results suggest that the increase in variability which ordinarily occurs with practice scores is not simply due to an increase with practice in the magnitude of the scores. The decrease in variability in the instance already mentioned (Test 5, Experiment I) is also against such an hypothesis. The suggestion is here offered that in further investigations an attempt should be made to determine the effect of variations in zeal and persistence of the subjects in long continued, monotonous practice experiments in order to ascertain the effect of such variations upon the variability of the group in the scores of the practiced tests.

From the preceding discussion, it should be clear that no sure conclusions concerning the effect of practice upon trait variability—the unevenness of scores of a particular individual in a number of tests—can be drawn from mere increase or decrease in test intercorrelations. Practice not only tends to increase group variability, but, as was the case in Experiment II if the conclusions reached in the preceding section are correct, may have an effect which is the equivalent of an increase in heterogeneity of the group—an increase of a sort which would increase correlation. An increase in correlation attributable to such a source would not necessarily indicate any decrease in trait variability. Suppose, for example, we took some second grade children and included their scores first in a group of scores made by children of their own ages and then in a group of scores made by children of widely different ages. In the second case the correlations would be greater; but the greater correlation in the second case obviously could not be taken to mean less trait variability on the part of the second grade children, for in the illustration it is assumed that the same second grade scores were used in both sets of calculations.

Standard errors of estimate.—The difficulty with which one is confronted in any attempt to measure trait variability is that of making scores in different tests comparable. Can they be made comparable simply by transforming them into standard scores and then reducing all values of $\sigma_{dis.}$ to unity? Obviously this can not be done, because the divisors which would need to be applied to the deviation scores vary at different stages of practice. For

example, in the case of Experiment II, to make $\sigma_{dis.}$ the same for Tests 1 and 4, the initial divisors would have to stand in the ratio 1.0 to 1.5, *i.e.*, $\frac{3.0}{4.6}$, while the final divisors, in the case of the same tests, would have to stand in the ratio of 1 to 2.3, *i.e.*, $\frac{3.7}{8.5}$.

A value which takes into account the change with practice in the ratio of the standard deviations of the correlated tests, better than does a correlation coefficient, is the estimate of scores in one test from scores in another test by means of the regression equation, $\overline{Y} = r_{xy}\frac{\sigma_y}{\sigma_x}(X - M_x) + My$. As may readily be observed, this equation takes into account the ratio of the standard deviations of the correlated tests. One could, perhaps, with some justification regard trait variability as something which increases, on the average for members of the group, whenever the errors of estimating the scores in one test from those of another in the case of a number of tests on the average show an increase. To determine these errors one could calculate each individual's score in one test from his score in a second test, take the difference between the actual and estimated scores, and finally take either the average of these differences or their standard deviation. The last mentioned value is probably the more satisfactory and it may be calculated by the simple formula for the standard error of estimate: $\sigma_{est.\ Y} = \sigma_y\sqrt{1 - r^2_{xy}}$. In the present study, the standard errors of estimate of scores in each test were calculated from each of the other tests, using first the data obtained from initial scores and then the data obtained from the final scores. Since scores in each of the seven tests may be predicted from those in each of the other six tests, forty-two standard errors of estimate were obtained, both for initial and final scores. By comparing these forty-two standard errors of estimate in the two cases, one may arrive at an answer to the question whether considerable practice in two tests increases the closeness with which scores in one test may be predicted from those in the other. The calculations reveal a statistically significant increase with practice in the average standard error of estimate in the case of Experiment

II, with the high school group. In the case of Experiment I, with the university group, however, there was no significant change in the average value of $\sigma_{est.}$. Practice caused a decrease in this average of only .063—scarcely greater than one-fourth the standard deviation of the decrease. Of the forty-two standard errors of estimate, twenty-nine increased and thirteen decreased. A number of the decreases, however, were rather pronounced, particularly in the case of Test 5, Digit Recognition, a test which was peculiarly affected by practice in that, while it showed an increase both as regards mean score and average correlation with other tests, it showed a decided decrease in the value $\sigma_{dis.}$. The data obtained concerning the standard errors of estimate are summarized in Table V.

TABLE V

AVERAGE STANDARD ERRORS OF ESTIMATE AS AFFECTED BY PRACTICE

	University (Exp. 1)	High School (Exp. II)
Mean initial $\sigma_{est.}$	6.899	6.658
$\sigma_{dis.}$ (N=42)	2.462	3.039
σ_M	.380	.469
Mean final $\sigma_{est.}$	6.836	7.828
$\sigma_{dis.}$ (N=42)	2.145	3.532
σ_M	.331	.545
Change in Means	—.063	1.170
$\sigma_{(M2-M1)}$	.233	.207
Significance ratio	.270	5.64
Ratio of means (final/initial)	.991	1.175
r (initial and final $\sigma_{est.}$'s)	.792	.927
Number of increases in $\sigma_{est.}$	29	33
Number of decreases in $\sigma_{est.}$	13	9

So far the discussion has dealt with the standard errors in estimating scores by means of fallible tests, that is, by tests of less than perfect reliability. Practice increases the reliability of the tests. This increase in reliability lowers the errors of estimate, both by decreasing the $\sigma_{dis.}$ and by increasing the value of r_{xy}. To be used as indices of true trait variability, that is, of trait variability if the traits were reliably measured, it is necessary to calculate the theoretical values of the standard errors of estimate for perfectly reliable measures. The true $\sigma_{dis.}$ is equal to the obtained $\sigma_{dis.}\sqrt{R}$, in which R stands for the reliabiltiy coefficient. The correction for imperfect reliability in r_{xy} is simply

the correction for attenuation. It follows that the value of $\sigma_{est.}$ with perfect reliability of tests becomes $\sigma_{y\ (true)}\sqrt{1-r^2_{\infty\infty}}$, in which $\sigma_{y(true)}$ is $\sigma_{dis.}$ of y after multiplication by the square root of the reliability coefficient of y, and $r_{\infty\infty}$ is the correlation between x and y after correction for attenuation. The results concerning the means, before and after practice, of the errors of the estimates after correction for test unreliability are given in Table VI.

A comparison of the results in Table VI with those in Table V shows that correction for unreliability of scores produces an

TABLE VI

AVERAGE STANDARD ERRORS OF ESTIMATE AFTER CORRECTION FOR TEST UNRELIABILITY

	University (Exp. I)	High School (Exp. II)
Mean initial $\sigma_{est.}$	6.167	5.944
$\sigma_{dis.}$ $(N=42)$	2.430	2.850
σ_M	.375	.440
Mean Final $\sigma_{est.}$	6.420	7.226
$\sigma_{dis.}$	2.171	3.428
σ_M	.335	.529
Change in Means (M_2-M_1)	+.253	1.282
$\sigma_{(M2-M1)}$	.244	.244
Significance ratio	1.037	5.254
Ratio of means (final/initial)	1.040	1.216
r (initial and final $\sigma_{est.}$'s)	.769	.890
Number of increases in $\sigma_{est.}$	33	33
Number of decreases in $\sigma_{est.}$	9	9

increase in the ratio of the average final $\sigma_{est.}$ to the average initial $\sigma_{est.}$. After correction for unreliability the data indicate that in both groups there is an increase with practice in the $\sigma_{est.}$ and in both groups thirty-three of the forty-two $\sigma_{est.}$'s were greater after practice than they were initially. It is interesting to note that the increase in the $\sigma_{est.}$ is much larger and much more significant in the case of the younger group. In fact, while the increase was over five times its standard deviation in the case of the high school group it was only 1.037 times its standard deviation in the case of the university group and therefore not significant.

3. *Standard deviations of standard scores.* The most direct measure of trait variability would seem to be the variation in the standard scores of a given individual. One could use either the mean variation of such scores, as did Woodrow (2), or their stand-

ard deviation, that is, the standard deviation of the individual's standard scores, as did Hull (1). Individual variability shown by the group, as measured by the $\sigma_{dis.}$, is of course always unity when calculated from standard scores (that is, after dividing each deviation from the mean by the ordinary $\sigma_{dis.}$). With this value may be compared the average of the standard deviations of the individual subjects, each calculated from the standard scores made by the same individual in the various tests. The average trait variability so measured is ordinarily less than unity, *i.e.*, less than the group variability.

It should be remembered, however, that standard scores are scores which indicate goodness, or standing in a group (though not in constant units unless the raw scores constitute constant units). The answer to the question, how good is a subject's score, however, clearly depends, when answered by means of a standard score, upon the group in which he is placed. Moreover, the change in a subject's standard score as he is placed first in one group and then in another would seldom be the same in the case of different tests. It follows that the difference in a given subject's standard scores in any pair of tests would ordinarily change as he were placed in first one group and then another. Conversely, a difference in standard scores in one group does not have the same significance, as regards trait variability, in one group that it has in another. In addition to the difficulty due to differences in heterogeneity is that due to the fact that the test reliabilities are not likely to be identical in the two compared groups; and a change in the reliability of the various tests would affect the variability in an individual's standard scores in those tests. Still another complication arises from the improbability that errors of measurement, *i.e.*, those which cause what is known as test unreliability, are distributed evenly over the whole range of scores. It seems certain that scores above the mean are more apt to contain positive errors and those below the mean negative errors than *vice versa*. These considerations have an important bearing on the present study because a given group of subjects is really two groups, psychologically speaking, when taken first at the beginning of practice and second at the end of practice. At

the end of practice, the group has changed as regards its heterogeneity, and differently as regards the various test-performances; and the test reliabilities have also all changed, and changed by varying amounts. It is clear from these considerations that no satisfactory picture of the change with practice in the true trait variability of the subjects could be obtained from the change in the average individual standard deviation in standard scores, when calculated first from initial and second from final scores. What, on first thought, seemed to be the most direct method of studying the problem was therefore finally abandoned because it appeared to present obstacles that were practically insuperable.

Summary.—Practice in seven memory tests was given to two groups of subjects, one consisting of forty-three university students (Experiment I) and the other composed of thirty-four tenth grade students (Experiment II). For each group the practice in each test consisted of twenty-five trials distributed over twenty-five days. The conclusions reached may be summarized as below.

(1) In both groups, practice produced an increase in correlation in the cases of the majority of the pairs of tests. Evidence was obtained which indicated that the increases in correlation in the case of the tenth grade group, which were marked and shown by twenty of the twenty-one pairs of tests, were due largely to an increase in heterogeneity of the group brought about to a considerable degree by failure on the part of some of the pupils to persist in putting forth their best efforts.

(2) In spite of the increases in test intercorrelations with practice, it was impossible to predict final test scores in one test, from those made in a different test, any better after practice than initially. In the case of the university group, practice did not significantly change the average of the forty-two standard errors of the estimates, but in the case of the high school group it produced a significant increase in this average value. When the standard errors of the estimates were corrected for test unreliability, increases in their average value as a result of practice were revealed on the part of both groups. In the case of the university group, the final average was 1.040 times the initial average;

and in the case of the high school students, the final average was 1.216 times the initial average. The increase in this average value of the standard errors of estimate was significant, whether or not the correction for unreliability was used, only in the case of the high school group. In both groups, however, thirty-three of the forty-two standard errors of estimate, after correction for test unreliability, were larger after practice than they were initially.

(3) The average values of the individual standard deviations of standard scores are not comparable as indices of trait variability because of the change with practice in the value of $\sigma_{dis.}$ by which each of an individual's seven scores (as deviations from the mean) is divided before calculating their standard deviation, and because of the effect of change in the reliability of the tests upon the variation in standard scores.

References

1. HULL, C. L. Variability in amount of different traits possessed by the individual. *J. educ. Psychol.*, 1927, **18**, 97–106.
2. WOODROW, H. Mental unevenness and brightness. *J. educ. Psychol.*, 1928, **19**, 289–302.
3. ——— The effect of practice on test intercorrelations. *J. educ. Psychol.*, 1938, **29**, 561–572.